FROM ORDERING EVALUATORS TO SELECTING ONE: WHAT A VALIDATION STATISTIC IS WORTH UNDER A LABEL BUDGET

Anonymous authors
Paper under double-blind review

ABSTRACT

Automatic evaluators are validated by pooled correlation with human labels over clustered items and used to choose inside a cluster. We measure what a validation statistic is worth when it selects one evaluator from a budget of labelled clusters, by the utility of the selected evaluator on clusters no rule used, with an estimand, a selection protocol and intervals whose coverage is measured inside worlds with a known or simulated truth. A statistic grouped on the decision's cluster orders evaluators better than pooling, and that better ordering is an unreliable guide to how much better an evaluator it selects, most visibly on machine translation, where large ordering advantages came with gains whose intervals included zero on every pool frozen in advance. Across public benchmarks for speech, summarization, story generation, machine translation and reward modelling, some of them read under protocols frozen in advance, grouped statistics usually order evaluator pairs better than pooled correlation, while the utility gain of the evaluator they select is uneven and its intervals often include zero. With few labelled clusters the sign of the gain differs by panel and rests on single tasks. A pre-registered, time-stamped manipulation of the candidates in a labelled cluster contradicted the account we had offered for that difference, which we withdraw, and on its replication set showed the grouped statistic's gain falling as candidates are removed. We recommend reporting, beside any validation statistic, the utility of the evaluator it selects with its interval, its leave-one-out and how the labels were spent.

1 INTRODUCTION

Automatic evaluators stand in for human judgment across language and speech generation: reward models and LLM judges choose among a policy's samples (Frick et al., 2025), metrics score summaries and rerank translations (Liu et al., 2023b; Freitag et al., 2022), and pronunciation scorers and MOS predictors grade speech (Gong et al., 2022; Huang et al., 2022a). Before one is used, it is chosen on human labels, often by the correlation of its outputs with the labels pooled over every item of a test set; reward-model benchmarks and, at the segment level, the WMT metrics task take a statistic inside the prompt or the source segment instead, where the decision is made (Lambert et al., 2025; Malik et al., 2026; Deutsch et al., 2023; Freitag et al., 2024). Which statistic predicts use has been asked across evaluators, by correlating statistics with outcomes (Wen et al., 2025; Frick et al., 2025) and by the evaluator a validation scheme selects (Guerdan et al., 2025; Collot et al., 2026).

Two choices are in play. A *candidate choice* is taken inside a cluster: which of a model's samples to return, which of a learner's sentences to correct. An *evaluator choice* is taken once, on labelled data: which evaluator to trust with the candidate choices. A pooled correlation can make the second choice badly at no visible cost. On the WMT22 en-ru MQM set, MATESE correlates better with the human scores over all translations (0.345 against 0.234 for YiSi-1), yet inside a source segment YiSi-1 orders more pairs of translations as the annotators did (65% against 57%). A statistic grouped on the segment repairs the ordering, but not the choice it feeds: over the set's whole panel, with half its segments labelled, the grouped mean names the worse metric of a pair 14.3 points less often than pooling, and the metric it selects is worth +1.0 $[-2.9, +5.0]$ points of best-of-n recovery, pooling's own pick being already within 3.0 points of the best metric.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

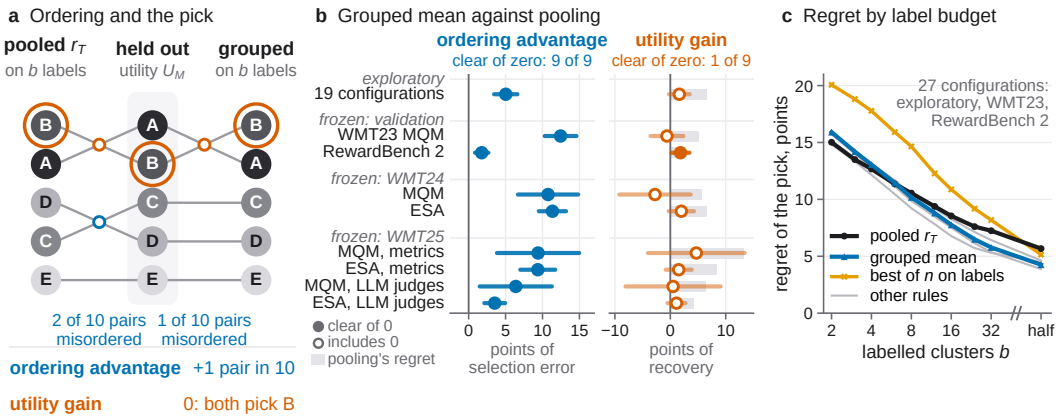


Figure 1: **Ordering advantage and utility gain of the grouped mean against pooling.** (a) Schematic: evaluators ranked on labelled clusters by pooled correlation and by the grouped mean, each joined to its rank by held-out utility; a crossing is a misordered pair, vermilion where it holds A, the held-out best. The grouped mean misorders one pair fewer, yet both name B: no gain. (b) At half the clusters, best of n: its ordering advantage and its pick's utility gain, each protocol's interval, on the exploratory set and the eight frozen sets of Table 1; grey, the regret of pooling's pick, which bounds any rule's gain. (c) Regret by label budget, best of n.

This paper asks what a statistic's better ordering of a panel is worth when it selects one evaluator from a budget of labelled clusters, measured by the utility of that evaluator on clusters no rule used, and it tests its readings under five protocols frozen before the statistics they grade were computed, the last four time-stamped by a third party. The contributions:

- **Ordering is not choosing (Figure 1, Table 1).** On panels of different systems, with half the clusters labelled, a statistic grouped on the decision's cluster orders evaluators better than pooling in the mean over a set's configurations, and the evaluator it selects is not reliably better than pooling's, the gain uneven across panels: under best of n its interval clears zero on one of the eight sets graded under frozen protocols and on both WMT25 ESA sets rerun with every rated system, and under the pairwise choice on two of the eight. The ordering advantage, predicted by the first frozen protocol, held on every set of different systems graded since; three readings of the sets read first were not confirmed by later protocols, and our explanation for why the grouped mean lost with eight labelled clusters on panels of four and six candidates a cluster, that a per-cluster statistic over so few values separates evaluators poorly, failed its frozen test and is withdrawn (Sections 5 to 7). That naming the best alternative and the cost of the one named are different criteria is textbook in ranking and selection (Hong et al., 2021); added here is how far apart they sit for evaluator selection, under tests fixed in advance.
- **What a set can show, and where the advantage falls (Table 1, Section 7).** Read after every protocol: a gain is bounded by pooling's regret, the recovery its pick leaves against the evaluation half's best evaluator, which with half the clusters is 2.6 to 13.2 points on the frozen sets and on two of them smaller than the gain the set detects with 80% power; with eight labelled clusters it is 9.1 to 18.1 points, and on six of the nine sets, the exploratory and the eight frozen, the interval excludes the grouped mean's closing more than 23% of it (Appendix P). On the 27 configurations read first, the 89% of decisive pairs that do not hold the evaluation half's best evaluator carry 92% of the ordering advantage at half, and the gain correlates with it there at +0.12.
- **An estimand and its calibration.** The expected paired utility, over half-splits by cluster, of the evaluators two rules select on one half, scored on the other: the held-out evaluation of a selection procedure (Cawley & Talbot, 2010; Xu et al., 2026) with two rules paired on the same held-out clusters, its intervals' coverage measured on the exploratory and first protocol's sets inside worlds with a known or simulated truth (Sections 4 and 5).
- **Choosing a judge from eight labelled prompts.** Tested by a fifth protocol on PPE's best-of-16 sets, with seven zero-shot LLM judges rating one model's samples as integers, the tie-calibrated acc^*_{eq} was not clearly better than pooling, +0.3 $[-0.1, +0.7]$ points, and eight labelled prompts rarely named the best of close judges. A second addendum, frozen after the single judges' ratings

and their mean on these sets had been read, found that averaging the three judges pooling ranks highest on the eight labels beats pooling's single pick on PPE and on both WMT25 sets, at three times the judge calls; three judges drawn at random, at the same calls and compared after the addenda, came within a point of that jury on PPE and fell well short of it on WMT25 (Section 8).

2 Related Work

Grouping. Splitting a correlation into between- and within-cluster parts is Robinson's ecological correlation (Robinson, 1950), restated by Hamaker (2024) and Tu et al. (2025), and the bound on how strongly a function of cluster identity can correlate with a label is Pearson's correlation ratio (Pearson, 1905); neither is ours. Summarization and translation have long debated which grouping a correlation should use (Deutsch et al., 2021; 2022; 2023; Agrawal et al., 2024; Lim et al., 2024), and a global and a within-source meta-evaluation rank LLM judges differently (Hu et al., 2025). Our r_W is the within-subject correlation of Bland & Altman (1995), the repeated-measures correlation of psychology (Bakdash & Marusich, 2017), and the PDP coefficient of DiIanni & Deutsch (2025) in a balanced crossed design, so it is measured here as a rule and not offered as a method. Gao et al. (2025) recommend pooling by ranking consistency and discriminative power, properties of a statistic, not of a decision, and Gao et al. (2026) score metrics by the labels they save in system comparison.

Validation against use. Machine translation added a tie-calibrated pairwise accuracy inside each source segment in 2023, beside pooled segment-level Pearson (Deutsch et al., 2023; Freitag et al., 2023), made it the only segment-level statistic in 2024 because metrics are used "to rank a set of translations for the same source segment", in minimum-Bayes-risk decoding and quality-estimation reranking (Freitag et al., 2024), and kept it (Lavie et al., 2025) (Appendix G); whether a better such statistic selects a better metric for choosing among translations was not measured, while a metric chosen for its correlation with human judgments has reranked into translations humans rated worse (Fernandes et al., 2022). An all-systems correlation is driven by systems far apart (Mathur et al., 2020; Xu et al., 2025), translation metrics are scored by the decisions they support (Kocmi et al., 2021; Perrella et al., 2024a), and the best of them now match a human reference in meta-evaluation (Proietti et al., 2025); for classifiers, an all-pairs ordering statistic can be an incoherent criterion for choosing between them (Hand, 2009).

Reward models and LLM judges. For reward models, statistics, per-prompt rank correlations among them, are compared by how well they predict downstream regret and preference across models (Wen et al., 2025; Frick et al., 2025), benchmark accuracy is often a weak guide to best-of-n and RLHF outcomes (Chen et al., 2024a; Kim et al., 2025; Razin et al., 2025), and a choice between LLM judges made on RewardBench can be the wrong one for reranking a generator's samples (Zhou et al., 2025). Landesberg (2026) decomposes an LLM judge's agreement into between- and within-prompt parts, finds coarse pointwise scores tied in two thirds of within-prompt pairs and best-of-2 pairwise judging regaining much of the lost signal, and recommends reporting within-prompt signal, tie rates and recovery beside global agreement. LLM judges have been compared by chance-corrected agreement, system rankings and reliability (Thakur et al., 2025; Gera et al., 2025; Norman et al., 2026); the mean of a judge's rating distribution beats its greedy integer (Wang et al., 2025), latent scores reduce the ties of Likert ratings (Girrbach et al., 2025), weak verifiers can be weighted into an ensemble with a few labelled queries (Saad-Falcon et al., 2025), judge designs have been tuned against human agreement (Salinas et al., 2025), and reward models have been read as metrics (Gehrmann, 2025). Three works score a validation statistic by the evaluator it selects: beside each reward-model statistic's correlation with regret, Wen et al. (2025) report the downstream rank of the model it ranks first, on synthetic gold labels; forced-choice agreement selects LLM judges up to 31% worse downstream (Guerdan et al., 2025); and balanced accuracy selects binary judges better over golden sets of varying size (Collot et al., 2026). None selects between a pooled and a grouped statistic from a few labelled clusters and scores the pick in payoff units on held-out clusters under frozen protocols; Wen et al. also split an annotation budget between prompts and responses. Where Landesberg (2026) measures the within-prompt signal, ties and recovery of LLM judges, we compare rules that select one evaluator, reward models and LLM judges among them, from a few labelled clusters by the held-out utility of the evaluator selected.

Selection with few labels. Active model selection spends a few labels to identify the best classifier or model (Okanovic et al., 2025; Kay et al., 2025), and prediction-powered estimators stretch a few human labels with an automatic evaluator (Boyeau et al., 2025; Dorner et al., 2025), treating labelled items as exchangeable rather than clustered, and a label budget can be spent on choosing and calibrating a panel of judges rather than one (Zhu et al., 2026; Li, 2026). What is added here is the held-out utility of the evaluator a statistic selects from a budget of labelled *clusters* (Section 4), how far it sits from the statistic's ordering advantage, and what the label budget does to both.

3 Cluster Structure in Evaluator Validation

Setup and decomposition. A benchmark provides N items; item i has a human label y_i, an evaluator output p_i and a cluster $c(i)$ of n_c items, a speaker, a document, a source segment or a prompt. A common validity statistic is the pooled correlation $r_T = \mathrm{corr}(p, y)$. For a vector v over items let $v_i^B = \bar{v}_{c(i)} - \bar{v}$ be its between-cluster part, $v_i^W = v_i - \bar{v}_{c(i)}$ its within-cluster part and $\eta_v^2 = \|v^B\|^2/\|v - \bar{v}\|^2$ the between-cluster share of its variation. The parts are orthogonal, so

$$r_T \;=\; \eta_p\,\eta_y\,r_B \;+\; \sqrt{1-\eta_p^2}\,\sqrt{1-\eta_y^2}\,r_W, \qquad r_B = \cos\!\big(p^B, y^B\big), \quad r_W = \cos\!\big(p^W, y^W\big), \tag{1}$$

exactly for any cluster sizes, a restatement of Robinson's ecological correlation (Section 2): the pooled coefficient mixes whether an evaluator ranks the clusters correctly, r_B, with whether it tracks differences inside one of them, r_W, weighted by the between-cluster shares of the labels and of the predictions. Only r_W bears on a choice among the items of one cluster. Given η_p and η_y, a pooled correlation bounds r_W only to an interval (Appendix B).

Statistics grouped on the cluster. r_W weights a cluster by how much the predictions vary in it and moves when one cluster's predictions are rescaled, which changes no decision (Appendix W). The paper therefore also scores the mean of the per-cluster Spearman correlations, grp, one coefficient per cluster with equal weight, and the pairwise accuracy inside clusters, a decision rather than a correlation; both are invariant to rescaling a cluster. Items of one cluster are dependent, so intervals resample clusters and are descriptive, cluster resampling under-covering when clusters are few (Cameron & Miller, 2015; Bowyer et al., 2025) and for r_T even when many (Appendix Z).

4 Choosing an Evaluator from Labelled Clusters

A *candidate pair* is two items of one cluster, and two decisions are scored on candidates: the pairwise choice, the share of candidate pairs whose labels differ that an evaluator orders as the labels do, and best of n, the share of the achievable label gain recovered by the item it ranks first in each cluster (Appendix H). A *selection rule* is computed on the *selection set*, b distinct labelled clusters every candidate of which carries a label, and selects one evaluator of a panel: the pooled Pearson coefficient r_T, the within-cluster coefficient r_W, the grouped mean grp, the payoff's own decision statistic on the labelled clusters, dec, and each benchmark's headline statistic, own (Appendix G).

The estimand. A practitioner selects one evaluator and obtains its utility on the clusters it is then used on, which are not the labelled ones. Scored on an evaluation set M drawn apart from the labelled clusters, the payoff of an evaluator e is a ratio of sums over M's clusters: for best of n, $U_M(e) = \sum_{c\in M} g_c(e) \big/ \sum_{c\in M} w_c$, with $g_c(e)$ the label gain e recovers in cluster c over the cluster's mean and w_c the gain recoverable there; for the pairwise choice, the concordant candidate pairs over the decisive ones. A rule's *utility* is the payoff of the evaluator it selects, and the estimand is the expected difference, over half-splits of a configuration's clusters (whole documents where clusters nest in them), between the utilities of two rules computed on one half and scored on the other, in the payoff's units (Algorithm 1). Given the corpus, the draws estimate it without bias, and the intervals, which resample the corpus's clusters, speak to corpora drawn like this one and not to other benchmarks; the paired difference is zero on every draw on which the two rules agree, so the estimand factors into the disagreement rate times the difference given disagreement; and both selected evaluators are scored on the same clusters, so M's own ranking is not taken as the truth. It is not the ratio of population totals; the gap is at most 5.7 points here (Appendix J, Table 9).

Selection error, and what an independent set does to it. The error of a rule is the share of evaluator pairs on which it selects the evaluator that is worse on M. Against an independent M that misorders a pair with its own chance q, an error e is observed as $e + q(1 - 2e)$, so every error moves toward one half and the average over pairs is reweighted and can reverse its sign (Appendix J); conclusions rest on utility. The two criteria also weigh the panel differently: the error counts each of the $K(K-1)/2$ pairs of K evaluators alike, while the utility of a pick turns only on the pairs that hold the evaluator named and on how far it falls short of the best, so a rule can repair many orderings among evaluators no rule would pick and gain nothing.

Draws and conventions. On PPE M is a fixed set of 192 further prompts a task that no rule uses; elsewhere every configuration's clusters are split in half 300 times (seed 20270915; PPE 500 draws), whole documents where clusters nest in them, a rule is computed on the selection half or a nested subset of b distinct clusters of it, and the evaluator it selects is scored on the evaluation half. One convention holds for every rule, target and budget: the pairs scored are those decisive on the target, the same for every rule; an undefined statistic or a tie in the rule is half an error; ties between evaluators are broken by one random priority per draw shared by every rule. Intervals are the estimate plus or minus 1.96 standard deviations of 200 bootstraps of each corpus with 300 splits each, clusters (documents on WMT) resampled jointly across a benchmark's configurations and every copy of a resampled cluster kept on one side of a split, the convention whose coverage Appendix AA measures; they carry the Monte Carlo error of the split average, 0.09 points on the exploratory mean at half the clusters and 0.15 at eight. One configuration's interval is wide: for the grouped mean against pooling, resampled from its own clusters, it excluded zero on 24.8% of corpora with eight labelled clusters and 15.1% at half (Table 73), so an interval that includes zero does not show a gain, or a loss, to be absent. Utility differences are in percentage points of the payoff, best-of-n recovery or pairwise accuracy; error differences are in percentage points of evaluator pairs. The primary analysis is the 19 configurations whose clusters are alternatives somebody chooses between, every numeric label of every benchmark whose release scores at least four evaluators on every item (Appendix H); configurations weigh equally, and against the validation sets of the frozen protocols they are the exploratory ones, the paper's protocol having been revised on them and the grouped mean having entered as a rule after pooled and within-cluster results on seven hand-chosen configurations had been read (Appendix I; terms: Appendix A).

5 What the Selected Evaluator Is Worth

Every rule is computed on half the clusters of each primary configuration and its pick scored on the other half (Table 5, Appendix H). Over the 19 configurations the grouped mean's evaluator has 1.6 points more best-of-n recovery than pooling's $[-0.2, +3.5]$ and 0.9 more pairwise accuracy $[+0.4, +1.4]$. The gain is uneven: under best of n it is positive on 11 configurations and negative on 4, no benchmark's interval clears zero, and HANNA relevance alone supplies 68% of the mean, which is +0.5 points without it (Tables 12 and 64); with the six benchmarks weighted equally instead of the configurations the mean is +2.1 $[-0.3, +4.4]$. Read as a bound, the exploratory interval excludes a gain above +3.5 points (+4.4 with the benchmarks weighted equally), while a single set's reaches +13.4 on WMT25 MQM (Table 1). Under worst of n, which pays for the candidate to flag or correct, the same picks remove 5.0 points of pooling's selection error at half and gain +3.2 $[+0.9, +5.5]$, against +1.6 under best of n, the two gains correlating at -0.01 across configurations (Appendix I; read after every protocol).

Ordering. On the exploratory set and on both validation sets, at half the clusters and on the same draws, the grouped mean names the worse evaluator of a pair less often than pooling, removing 18% of pooling's error under best of n and 40% under the pairwise choice on the exploratory set, with intervals clear of zero where the utility gain's is not (Tables 1 and 43; the first frozen protocol predicted it, below). The scope is a set mean on panels of different systems with more than a handful of clusters: under best of n the advantage is not positive at half on 7 of the 27 configurations, and it is negative on RewardBench 2 at every budget up to 48 prompts and on PRM800K at every budget.

What pooling leaves, and what a set can show. Beside each gain Table 1 sets pooling's regret, the points of recovery by which its pick falls short of the evaluation half's best evaluator on the same draw, and the gain the set detects with 80% power, 2.8 standard deviations of its bootstrap, both read

Table 1: **The grouped mean against pooling at half the clusters.** Best of n, on the exploratory set and on every set of different systems whose gain at half a frozen protocol graded. Adv.: the points of selection error the grouped mean removes against pooling, every interval clear of zero but RewardBench 2's without the authors' models, $[-1.1, +0.4]$; gain: the points of best-of-n recovery its pick earns over pooling's, with its interval (300 splits a replicate; the frozen 40 on WMT24 and WMT25); Reg.: the points by which pooling's pick falls short of the evaluation half's best evaluator; Det.: the gain the set detects with 80% power; these two read after every protocol (Appendix P). RewardBench 2 without the authors' models was fixed by the first protocol and not graded; the WMT25 rows froze the eight systems of a language pair last by name, the every-system rows keep all 18 to 20, rerun after the grading (Table 52).

Set	Adv.	Gain	Reg.	Det.	Set	Adv.	Gain	Reg.	Det.
exploratory, read before any protocol					*fourth protocol, WMT25, primary metrics*				
19 configurations, six benchmarks	+5.0	+1.6 [−0.2, +3.5]	6.6	2.7	MQM, eight systems frozen	+9.4	+4.7 [−4.0, +13.4]	13.2	12.5
first protocol					every system	+10.3	+5.1 [−4.6, +14.7]	11.3	13.8
WMT23 MQM	+12.4	−0.6 [−3.6, +2.4]	5.2	4.2	ESA, eight systems frozen	+9.4	+1.5 [−0.8, +3.9]	8.3	3.4
RewardBench 2	+1.8	+1.8 [+0.2, +3.5]	2.6	2.3	every system	+11.8	+3.1 [+0.8, +5.5]	10.2	3.3
without the authors' models	−0.3	+0.6 [−0.7, +2.0]	1.4	1.9	*fourth protocol, WMT25, thirteen LLM judges*				
third protocol, WMT24, primary metrics					MQM, eight systems frozen	+6.4	+0.5 [−8.1, +9.1]	6.3	12.3
MQM	+10.8	−2.8 [−9.2, +3.7]	5.6	9.2	every system	+7.8	−1.2 [−7.7, +5.3]	3.4	9.3
ESA	+11.3	+2.0 [−0.3, +4.3]	6.5	3.3	ESA, eight systems frozen	+3.5	+1.1 [−0.5, +2.7]	4.2	2.3
					every system	+2.7	+3.4 [+2.0, +4.8]	5.2	2.0

after every protocol (Appendix P). A rule's gain is pooling's regret less its own, so the regret bounds it, loosely, as the evaluation half's best is named with that half's noise. At half the grouped mean closes 25% of pooling's 6.6 points on the exploratory set; on WMT24 MQM and WMT25 MQM with the judges the detectable gains, 9.2 and 12.3 points, exceed pooling's whole regret, 5.6 and 6.3, the power against a gain that large being 40% and 30%; on six of the nine sets, the exploratory and the eight frozen, the interval's upper end lies below the regret, bounding the share closed at 46% to 65%. With eight labelled clusters pooling leaves more, 9.6 points on the exploratory set and 9.1 to 18.1 on the frozen sets, and on six of the nine the interval excludes the grouped mean's closing more than 23% of it (on RewardBench 2 it loses), the three MQM sets graded after WMT23 allowing 42% to 66%. No threshold was fixed in advance, so these are bounds and not tests of equivalence.

The within-cluster coefficient. Read after every protocol, no interval of r_W's pick against pooling's under best of n lies wholly below zero, on 24 set means and 140 configuration cells with eight labelled clusters or half, where each other alternative's does on 5 to 13 cells (Appendix Q). That counts intervals and is no test of non-inferiority: its worst set mean is -2.7 points, it lost on 3 of 222 further cells and with two candidates a cluster, and it departs from pooling least, on 47% of draws with eight labelled clusters against 63% for the grouped mean, though given a departure its pick gained +3.5 points on the exploratory set against the grouped mean's +1.9.

The first frozen protocol. It predicted that at half the clusters the grouped mean's evaluator gains over pooling's (H1) and that it orders evaluator pairs better than pooling (H3), with three further readings (H2, H4, H5; Appendix S); it, its grading rule and its code were hashed (`prereg/`, 2026-09-15 20:50:36 UTC) before either validation set was read: *WMT23 MQM*, en-de, he-en and zh-en with 32 primary metrics, and *RewardBench 2* (Malik et al., 2026), 174 reward models on five subsets of four completions a prompt, one of them correct, so that its labels have no between-prompt part and Equation 1 makes pooling $r_T = \sqrt{1-\eta_p^2}\, r_W$; 132 of the models are the benchmark authors' training runs (Appendix S). On WMT23 the grouped mean orders evaluator pairs far better than pooling, 12.4 points fewer wrong under best of n, while the metric it selects is worth -0.6 points of recovery $[-3.6, +2.4]$ (H3 supported, H1 not supported). On RewardBench 2 both are supported, the gain carried by the math subset and both resting on the benchmark authors' checkpoints: without them H3 is not supported, and with one model per family the gain is +0.4 points, its interval clear of zero on 0 of 20 draws (Appendix S). H2 and H5 are not supported on WMT23 (Appendix R).

6 Labels, Candidates and Samples of One Model

The label budget. With eight labelled clusters (Figure 6, Appendix K) the grouped mean's evaluator has 1.2 points more best-of-n recovery than pooling's $[+0.1, +2.2]$, clear of zero by less than the Monte Carlo error of its split average. From eight clusters to half its ordering advantage doubles while the gain barely moves (Table 35).

Samples of one model. Choosing among one model's samples for one prompt no released panel covered, so two checks were built under designs written before any evaluator was scored (Appendices L and M): GPT-4 solutions of PRM800K (Lightman et al., 2024) with 5 public reward models, and the best-of-K sets of PPE (Frick et al., 2025) with five reward models, a length baseline and 192 further prompts a task as a fixed evaluation set. With eight labelled problems on PRM800K the grouped mean's pick gives up 7.3 points of recovery against pooling's $[-9.6, -5.0]$ under the main run's seed and 4.3 to 6.1 under twelve others; with eight labelled prompts on PPE, up to 128 labelled responses, the within-cluster coefficient's pick gains +1.6 points of best-of-16 accuracy $[+0.7, +2.5]$, pool and measurement set resampled, carried by one task of five, MBPP-Plus (Table 21).

The candidates in a labelled cluster, moved. The grouped mean lost to pooling with eight labelled units on the two panels whose clusters hold four and six candidates and gained on the one whose prompts hold sixteen, suggesting that a per-cluster statistic over so few values separates a panel poorly. The second protocol kept m randomly drawn candidates in every labelled cluster and predicted that the grouped mean's gain falls as candidates are removed and turns negative at four on PPE (time-stamped 2026-09-18 22:21:58 UTC, before any rule was computed on a subsampled cluster; Appendix N). On PPE the gain stayed positive at four responses, +0.9 $[+0.1, +1.7]$ points, a contradicted prediction, so the explanation is withdrawn, as the protocol said. On the exploratory configurations it falls as predicted, but at equal labels, read after the fact, four candidates in place of eight cost it only +0.3 to +0.7 points (Tables 27 to 29): the fall is mostly the removed candidates' labels.

7 What an Ordering Advantage Buys

Read after the fact. Across the 27 configurations read first (Figure 8, Appendix O; descriptive, as they share benchmarks), the grouped mean's ordering advantage and the utility gain of its pick correlate at 0.68 with eight labelled clusters and 0.19 at half under best of n. Most of the advantage falls on pairs that do not hold the evaluation half's best evaluator, about in proportion to their number: at half they are 89% of the decisive pairs and carry 92% of it, and the gain correlates with the per-pair advantage on the pairs that do at 0.47, partly by construction, and on the rest at +0.12; a simulation that distorts one rule's limit ordering away from the top three reproduces the loosening (Appendix O). In regret (Figure 1c), with eight labelled clusters the grouped mean's pick is as good as pooling's with 9.1, and best of n on the labelled clusters leaves the most at every budget up to 32, a reading that held weakly when the fourth protocol tested it on WMT25's primary metrics and not when the fifth tested it with LLM judges on PPE.

Frozen on WMT24. Among the sets read first, ordering and choice parted most on machine translation, with advantages of 9 to 19 points at half and a mean gain near zero. The third protocol froze it before any WMT24 metric score was read (Appendix T): an advantage of at least 8 points (M1) and a gain below the line fitted outside translation (M2). It was confirmed on ESA and neither confirmed nor failed on MQM, where M1 under best of n was consistent rather than supported (Table 1, Appendix R), but on both sets M2 could not fail, pooling's regret lying below the line.

Frozen on WMT25, with thirteen LLM judges. The fourth protocol was frozen before any WMT25 data file was read, though the year's two findings papers, with the metrics task's published segment-level rankings, had been (Appendix U). Two panels scored fifteen WMT25 language pairs: the primary metrics, and thirteen general-purpose LLMs prompted zero-shot by the organisers (Lavie et al., 2025). The ordering advantage (N1) is confirmed with both panels. Whether a configuration's gain tracks its advantage (N2) is not settled: with the primaries a configuration's gain varies more across its own bootstrap than across the fifteen configurations, so the classical ceiling on that correlation is 0.00 with eight segments, and advantage and gain share noise within a configuration

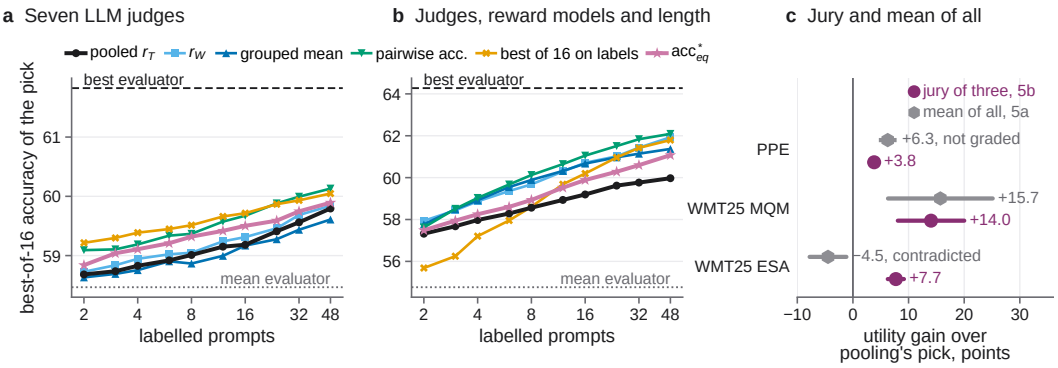


Figure 2: **The fifth protocol and its two addenda.** (a, b) On PPE, best-of-16 accuracy on the fixed measurement prompts of the judge each rule names, by the labelled prompts it reads, the five sets' mean, for the seven LLM judges and with the reward models and the length baseline. (c) With eight labelled units, the gain over pooling's pick of the jury of three (5b) and of the mean of all judges (5a; dashed on PPE, read after the grading), each protocol's interval. Grades: Table 42; intervals: Tables 57, 60 and 61.

(replicates correlating at 0.44), which alone can produce the supported 0.58 with eight segments. Neither set confirmed the third protocol's reading, and with the primaries the MQM gain at half, +4.7 points, failed M3's test of a small gain. The reading registered on 19 September, before this test, that large ordering advantages come with no gain, holds for the frozen ESA set mean and not for the MQM set, across the ESA language pairs or with every system kept (below); and M2's line owes its height to one configuration, HANNA relevance (Appendix U).

The pool and the judges. Rerun after the grading (Table 52), the advantage holds and the gain moves with the pool. The frozen adapter drops the system with the fewest rated segments, ties broken by name, until 300 segments are rated for every system kept or eight remain; WMT25 rates every system on the same segments, so it kept the eight systems last by name. With all 18 to 20 systems of a language pair both ESA gains at half clear zero, +3.1 $[+0.8, +5.5]$ points with the primaries and +3.4 $[+2.0, +4.8]$ with the judges (Table 1), the judges' only while the systems of the 3 to 7 judges that score their own translations stay in the pools. The judges score coarsely and often tie at the top (Landesberg, 2026); among them, described and not graded, the tie-calibrated acc^*_{eq} (Deutsch et al., 2023) picked a better judge than pooling with eight labelled segments in every rerun, +5.6 $[+3.5, +7.7]$ points on ESA, which the fifth protocol tested on one model's samples (Section 8).

8 Choosing an LLM Judge for Best of 16

The fifth protocol asked which LLM judge, read as zero-shot pointwise ratings, to trust with best of 16 over one model's samples given a few labelled prompts. Before any judge had rated a response of the pool or the measurement set, it froze twelve general-purpose LLMs, each rating one response from 1 to 10, and five predictions for the five sets' mean (Table 42, Appendix R), the main one, L2, that with eight labelled prompts acc^*_{eq}'s pick would beat pooling's, every rule paid the best-of-16 accuracy of the judge it names on the 192 fixed prompts of a set (time-stamped 2026-09-25 15:11:06 UTC, PPE's labels and the reward models' scores having been read for Section 6; Appendix V). A spent API account cut the panel to the seven judges that had rated at least 95% of both sets, five with open weights, by a deviation recorded before any rating was read.

No prediction that could fail was supported on the seven judges (Figure 2a). acc^*_{eq}'s pick gained +0.3 $[-0.1, +0.7]$ points over pooling's (L2 consistent), nearly all on MATH; the grouped mean did not clearly order judge pairs better at half the pool (L1 consistent); best of 16 on the labelled prompts did not pick worse than pooling (L5 not supported); and L4, supported, could not have failed, pooling's regret at half, 2.0 points, leaving no room for a gain of three. Eight labelled prompts say little about seven close judges: every rule named the best judge of a set on 14% to 20% of draws, against 14% for a random pick, pooling's pick falling 2.8 points short of each set's best judge on average. The

deviation's cut moves two grades: L1's estimate turns negative on the fully rated prompts, and with all twelve judges best of 16 on the labels picks worse than pooling, as L5 predicted (Appendix V). With the reward models and the length baseline, pooling's pick fell 5.7 points short; the grouped mean ordered evaluator pairs better at half (L1 supported), and the plain within-prompt statistics recovered 1.1 to 1.6 points with eight labelled prompts, mostly on MBPP-Plus and IFEval, while acc^*_{eq}'s was still not clearly better (L2 consistent).

By the protocol's rule, choosing a coarse LLM judge by acc^*_{eq} rather than pooled correlation with eight labelled prompts, up to 128 labelled responses, is neither confirmed nor failed as advice. Read after the grading, the label-free mean of the seven judges' ratings beat the best single judge of each set by 3.4 points $[1.2, 4.6]$ and pooling's pick with eight labels by 6.3, as a jury beat a GPT-4 judge elsewhere (Verga et al., 2024). A frozen addendum (5a in Table 42) put the mean to WMT25's thirteen judges: against pooling's pick with eight labelled segments it gained +15.7 $[+6.3, +25.1]$ points on MQM and -4.5 $[-7.8, -1.2]$ on ESA (contradicted), where on en-ar and en-bho most judges, and so the mean, pick worse than chance. A second (5b; Figure 2c) spent the eight labels on choosing whom to average: the three judges pooling ranks highest, averaged, beat its single pick by +3.8 $[+3.0, +4.6]$ points on PPE, +14.0 $[+8.1, +20.0]$ on MQM and +7.7 $[+6.3, +9.2]$ on ESA, all supported, and beat the mean of all judges on ESA alone. Three judges drawn at random, at the same three calls and compared after the addenda (Table 62), came within a point of the jury on PPE, which led them by +0.2 $[-0.3, +0.7]$ points; on WMT25 the jury led them by +8.8 $[+2.8, +14.9]$ on MQM and +19.8 $[+17.9, +21.8]$ on ESA. By its rule the text advises, with eight labels, averaging the three judges pooled correlation ranks highest rather than using the one it ranks highest, at three judge calls a response instead of one; 5b was frozen after every single judge's ratings and their mean on these sets had been read, so fresh prompts, judges and years have not yet tested it.

9 Limitations and Conclusion

On panels of different systems, with half the clusters labelled, a statistic grouped on the decision's cluster orders evaluators better than pooling in the mean over a set's configurations, and the evaluator it selects is not reliably better than pooling's, the gain uneven across panels: under best of n its interval clears zero on one of the eight sets graded under frozen protocols and on both WMT25 ESA sets rerun with every rated system, and under the pairwise choice on two of the eight. At half the clusters the grouped mean's evaluator is worth +1.6 $[-0.2, +3.5]$ points of best-of-n recovery on the exploratory set, 25% of what pooling's pick leaves, and on the frozen sets what Table 1 gives, the LLM judges' WMT25 ESA gain with every system kept clearing zero only while they score their own systems' translations. Set against pooling's regret, which caps any rule's gain, the intervals bound the gain at a small share of it on most of the nine sets, the exploratory and the eight frozen, with eight labelled clusters; at half an interval's upper end lies below the regret on six of them and above it on RewardBench 2 and both WMT25 MQM sets, and on two MQM sets even a gain of the whole regret would have been shown at less than even odds. Three readings of the sets read first were not confirmed by later protocols, and our explanation for the sign of the gain with eight labelled units failed its frozen test and is withdrawn. The five protocols and the fifth's two addenda gave 135 grades (Table 42, Appendix R); each protocol was written after the grades before it had been read, so they are sequential tests of earlier readings on new data, not one confirmatory family, and carry no multiplicity correction.

The limits: text and speech in English only outside translation; panels of different systems except on PRM800K and PPE, some of five evaluators; one year, WMT25, of new data for the clauses of Section 7, its frozen pools the eight systems last by name, its MQM sets two language pairs and its ESA label one rating; seven LLM judges in the fifth protocol, five of one family, all served by one provider and read as greedy integers from one prompt, their panel cut by a spent API account; intervals outside the calibration on PRM800K, PPE, WMT24 and WMT25 (Appendix AA). For a practitioner, an ordering advantage is not a selection gain, and an interval that includes zero bounds the gain only where its upper end falls below pooling's regret. Eight labels rarely named the best of close evaluators; averaging the three LLM judges they rank highest beat choosing one, at three times the judge calls, under an addendum frozen after the ratings had been read, and on PPE three judges drawn at random came within a point of it. A benchmark can report, beside any validation statistic, the utility of the evaluator it selects with its interval, the regret pooling leaves and the gain the set can detect, its leave-one-out and how the labels were spent (Appendix D).

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing the text and for generating synthetic datasets. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work reuses public benchmarks and collects no new data from human subjects, so no ethics-board review was sought; the paragraph below says what we think a reviewer should weigh instead. All human labels analysed here were released by their original authors under the licences Appendix Y lists, and we report only aggregate statistics over them; where a release leaves its text to third parties' terms or asks that it not be shown online, as PPE does for its models' responses and GPQA for its questions, the supplement ships no text, only hashes and the scripts that rebuild it. Rater and listener identifiers are used solely to estimate reliability and are never linked to any other information. One benchmark, BVCC, distributes part of its audio only through the original challenge archives, which we fetched with the release's own scripts and do not redistribute.

Appendix X sends audio to two commercial interfaces and benchmark text to a third, and a public licence does not make a recording cease to carry information about the person who spoke. speechocean762 is an openly licensed corpus of second-language learners whose recordings are already distributed worldwide, many of them children: of the 125 test speakers whose recordings we sent, 64 are under 18 and 52 are 12 or under. We send the released file and the released reference text and nothing else: no speaker identifier, no label, no age or group beyond the one field iFlytek's interface requires, and no history (Appendix X). What we cannot report is what the vendors then do. We used the ordinary paid and trial tiers rather than any enterprise agreement, so we did not negotiate a retention or no-training term, we did not pin a processing region, and we have not obtained a written confirmation from any of the three that their terms permit publishing per-item results measured through their interfaces. We name that gap rather than assert compliance we did not verify. A benchmark holder repeating this audit on a corpus whose speakers are identifiable, or one whose licence is narrower than speechocean762's, should settle those terms first, and should consider whether transmitting the voices onward is within the consent the original collection obtained, which a redistribution licence does not by itself establish. The fifth protocol sent PPE's released prompts and responses, text without personal data, to one commercial API for ratings under its ordinary terms, a pilot's requests to a second having been refused (Appendix V); the responses are outputs of Claude 3 Haiku, Llama-3-8B-Instruct, GPT-4o-mini and Gemma-2-9B-it, whose providers' terms we did not check for passing them to another provider's API or for publishing per-item ratings of them; the caveat about publishing per-item results applies to those ratings too.

We name the services and report the same statistics for them as for every other evaluator. Because many of the speakers are children and the terms are unconfirmed, the supplement does not include the two speech services' per-utterance scores, only the aggregates computed from them and the scripts; the judge's ratings of public summaries and the fifth protocol's ratings, from which its grades are recomputed, are included. The purpose is to measure what a pooled correlation certifies, not to rank vendors, and the numbers are what those interfaces returned on the dates recorded in Appendix X rather than a property of a fixed artifact. The analyses can lower published correlation numbers for existing evaluators, which is their purpose, and we report reproduced results for every system we ran rather than relying on numbers we could not check.

REPRODUCIBILITY STATEMENT

All benchmarks used in this paper are publicly available, and Appendix Y lists sources, versions, and preprocessing. The supplementary material contains the anonymized toolkit as an installable package with its simulation tests, the normalisation scripts that turn each raw benchmark into the analysis tables, the scripts that produce every table and figure, the per-item predictions of all reproduced evaluators and of the commercial LLM judge, the two speech services' per-utterance scores withheld as the ethics statement says, and one card per configuration of the census. One label array is withheld because its licence forbids redistribution, with the recipe to rebuild it from the source. The two same-model experiments of Appendices L and M ship as the design documents written before scoring, a timing that rests on the authors' word and the files' times, the item tables

built from PRM800K and, without their text, from the PPE sets, with the scripts that rebuild them byte for byte, the per-solution and per-response scores of every reward model and the analysis scripts, so both can be rerun without a GPU, the PPE set with its fixed evaluation set and the addendum that fixed it; the choice of an evaluator ships as one engine that computes every rule, every target and every budget on the same draws, with its per-configuration results, the corpus bootstraps, the simulation of the measurement targets and the audit of undefined statistics at every budget; the validation of Section 5 ships as the protocol and the hash record written before its data were read (`prereg/`), the adapters for WMT23 and RewardBench 2 written afterwards, the fetch record of the RewardBench 2 files and the graded report; the manipulation of Section 6 as its protocol, the hash record of the protocol and of every script it runs, the RFC 3161 time-stamp token of that record with the authority's certificate chain, the graded report, its addendum with a record and a token of its own, and the scripts of everything read after the fact beside it (the fixed-label arm, the same-systems variant, the other seeds, the coverage of its intervals and the relation between ordering advantage and utility gain by budget, by task family and by where in the panel the advantage sits, `scripts/decision/ordering_by_budget.py`, `ordering_by_budget_family.py`, `ordering_by_budget_parts.py` and `r15_thesis_checks.py`, with the simulation `scripts/sim/ordering_gain_mechanism.py`); the third protocol of Section 7 as its text (`prereg/PROTOCOL_mt_parting.md`), its hash record and time-stamp token, the adapter, the runner and the grader it froze (`scripts/normalize/wmt24.py`, `scripts/decision/run_mt_parting.py`, `mt_parting_report.py`), the adapter's cards and the graded report; the analyses added after the third protocol was graded, all read after the fact, as `scripts/decision/r16_headroom.py` (the regrets by budget and their exchange rate in labels), `wmt24_winners_r16.py` (the metric each rule names on WMT24, a replay of the frozen analysis), `candidate_alloc_all8_r16.py` and `candidate_seeds_replay_r16.py`, with `scripts/figures/fig_parting.py` for Figure 8 and `fig_teaser.py` for Figure 1, every figure drawn in the style of `figstyle.py`; the fourth protocol as its text (`prereg/PROTOCOL_wmt25.md`), its hash record and token, its two deviations, each with a record and token of its own, the adapters, the runners and the graders it froze or its deviations replaced (`scripts/normalize/wmt25_llm.py`, `scripts/decision/run_wmt25.py`, `run_wmt25_d1.py`, `wmt25_report.py`, `mt_parting_report_d2.py`), the adapters' cards and the graded reports, and, read after the grading, the reruns outside the protocol (`scripts/normalize/wmt25_variants.py`, `scripts/decision/run_wmt25_variants.py`) and the descriptive facts of Table 56 and of the correlations' noise (`wmt25_described_r17.py`, `wmt25_noise_ceiling_r17.py`), the reruns' per-configuration outputs left out for the size limit and regenerated by their runner; the fifth protocol as its text (`prereg/PROTOCOL_ppe_llm.md`), its two hash records with their tokens and its deviation, the scoring script with the judge prompt it froze, the pilot, every judge's ratings (reduced to the item and the rating for the size limit, the full replies hashed in the time-stamped results manifest `prereg/RESULTS_manifest_r18`), the runner and the grader (`scripts/same_model/ppe_llm_run.py`, `ppe_llm_report.py`), their bootstrap replicates and the graded report, and its addendum on WMT25 as its text (`prereg/PROTOCOL_ensemble_wmt25.md`), its hash record and token, its script (`scripts/decision/ensemble_wmt25_r18.py`) and results, hashed in a second time-stamped manifest (`prereg/RESULTS_manifest_r18b`), and its second addendum likewise (`prereg/PROTOCOL_jury.md`, `scripts/decision/jury_r18.py`, `prereg/RESULTS_manifest_r18c`), with its comparison against three judges drawn at random, read after it (`scripts/decision/jury_baselines_r22.py`); the tally of every rule's losses against pooling over the runs above, read after every protocol (`scripts/decision/rule_safety_r22.py`); pooling's regret and the gain each set can detect, read after every protocol (`scripts/decision/headroom_power_r23.py`); and, since the shipped scripts have their machine paths rewritten, every file the freeze records name byte for byte as hashed, with a script that checks all of them (`prereg/frozen_files.tar.gz`, `prereg/verify_frozen.py`); a change log lists every correction with the draft it belongs to. Proofs are in Appendix B and estimator details in Appendix D. The verification of Proposition 4, including the exhaustive enumeration over weak orders and a counterexample to a weaker form of the bound, is in `scripts/sim/per_row_bound.py`.

REFERENCES

Sweta Agrawal, António Farinhas, Ricardo Rei, and Andre Martins. Can automatic metrics assess high-quality translations? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14491–14502, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.802. URL https://aclanthology.org/2024.emnlp-main.802/.

Jonathan Z. Bakdash and Laura R. Marusich. Repeated measures correlation. *Frontiers in Psychology*, 8:456, 2017. doi: 10.3389/fpsyg.2017.00456.

Andrew Bell and Kelvyn Jones. Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods*, 3(1):133–153, 2015. doi: 10.1017/psrm.2014.7.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. Re-evaluating evaluation in text summarization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9347–9359, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.751. URL https://aclanthology.org/2020.emnlp-main.751/.

J. Martin Bland and Douglas G. Altman. Statistics notes: Calculating correlation coefficients with repeated observations: Part 1—correlation within subjects. *BMJ*, 310(6977):446, 1995. doi: 10.1136/bmj.310.6977.446.

Sam Bowyer, Laurence Aitchison, and Desi R. Ivanova. Position: Don't use the CLT in LLM evals with fewer than a few hundred datapoints. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 81143–81184, 2025. URL https://proceedings.mlr.press/v267/bowyer25a.html.

Pierre Boyeau, Anastasios Nikolas Angelopoulos, Tianle Li, Nir Yosef, Jitendra Malik, and Michael I. Jordan. AutoEval done right: Using synthetic data for model evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5276–5290, 2025. URL https://proceedings.mlr.press/v267/boyeau25a.html.

William Brown. Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3):296–322, 1910. doi: 10.1111/j.2044-8295.1910.tb00207.x.

A. Colin Cameron and Douglas L. Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015. doi: 10.3368/jhr.50.2.317.

Joshua Camp, Tom Kenter, Lev Finkelstein, and Rob Clark. MOS vs. AB: Evaluating text-to-speech systems reliably using clustered standard errors. In *Proc. Interspeech 2023*, pp. 1090–1094, 2023. doi: 10.21437/Interspeech.2023-2014.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9263–9274, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.745. URL https://aclanthology.org/2020.emnlp-main.745/.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010. URL http://jmlr.org/papers/v11/cawley10a.html.

Fu-An Chao and Berlin Chen. Towards efficient and multifaceted computer-assisted pronunciation training leveraging hierarchical selective state space model and decoupled cross-entropy loss. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1947–1961, 2025. doi: 10.18653/v1/2025.naacl-long.98.

Fu-An Chao, Tien-Hong Lo, Tzu-I Wu, Yao-Ting Sung, and Berlin Chen. A hierarchical context-aware modeling approach for multi-aspect and multi-granular pronunciation assessment. In *Proc. Interspeech 2023*, pp. 974–978, 2023. doi: 10.21437/Interspeech.2023-550.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113. URL https://arxiv.org/abs/2110.13900.

Yanjun Chen, Dawei Zhu, Yirong Sun, Xinghao Chen, Wei Zhang, and Xiaoyu Shen. The accuracy paradox in RLHF: When better reward models don't yield better language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2980–2989, Miami, Florida, USA, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.174. URL https://aclanthology.org/2024.emnlp-main.174/.

Yu-Wen Chen, Zhou Yu, and Julia Hirschberg. MultiPA: A multi-task speech pronunciation assessment model for open response scenarios. In *Interspeech 2024*, pp. 297–301, 2024b. doi: 10.21437/Interspeech.2024-123.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 5794–5836, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.509/.

Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. Do language models enjoy their own stories? Prompting large language models for automatic story evaluation. *Transactions of the Association for Computational Linguistics*, 12:1122–1142, 2024. doi: 10.1162/tacl_a_00689. URL https://aclanthology.org/2024.tacl-1.62/.

Michael Chinen, Jan Skoglund, Chandan K. A. Reddy, Alessandro Ragano, and Andrew Hines. Using rater and system metadata to explain variance in the VoiceMOS challenge 2022 dataset. In *Proc. Interspeech 2022*, pp. 4531–4535, 2022. doi: 10.21437/Interspeech.2022-799.

Sangwoo Cho, Kushal Chawla, Pengshan Cai, Zefang Liu, Chenyang Zhu, Shi-Xiong Zhang, and Sambit Sahu. Ask, don't judge: Binary questions for interpretable LLM evaluation and self-improvement, 2026. URL https://arxiv.org/abs/2606.27226.

Paul S. Clarke and Annalivia Polselli. Double machine learning for static panel models with fixed effects. *The Econometrics Journal*, 29(1):69–86, 2026. doi: 10.1093/ectj/utaf011. Advance access 2025.

Stephane Collot, Colin Fraser, Justin Zhao, William F. Shen, Timon Willi, and Ilias Leontiadis. Balanced accuracy: The right metric for evaluating LLM judges – explained through Youden's J statistic. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 927–936, Rabat, Morocco, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-industry.69.

Erica Cooper and Junichi Yamagishi. How do voices from past speech synthesis challenges compare today? In *Proc. 11th ISCA Speech Synthesis Workshop*, pp. 183–188, 2021. doi: 10.21437/SSW.2021-32. URL https://arxiv.org/abs/2105.02373.

Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi. Generalization ability of MOS prediction networks. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8442–8446, 2022. doi: 10.1109/ICASSP43922.2022.9746395.

Fredrik Cumlin. Rho-Perfect: Correlation ceiling for subjective evaluation datasets. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 15142–15146, 2026. doi: 10.1109/ICASSP55912.2026.11464989. URL https://arxiv.org/abs/2602.08552.

Daniel Deutsch, Rotem Dror, and Dan Roth. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146, 2021. doi: 10.1162/tacl_a_00417. URL https://aclanthology.org/2021.tacl-1.67/.

Daniel Deutsch, Rotem Dror, and Dan Roth. Re-examining system-level correlations of automatic summarization evaluation metrics. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 6038–6052, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.442. URL https://aclanthology.org/2022.naacl-main.442/.

Daniel Deutsch, George Foster, and Markus Freitag. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12914–12929, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.798. URL https://aclanthology.org/2023.emnlp-main.798/.

Colten DiIanni and Daniel Deutsch. Don't sweat the small stuff: Segment-level meta-evaluation based on pairwise difference correlation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25062–25070, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1273. URL https://aclanthology.org/2025.emnlp-main.1273/.

Florian E. Dorner, Vivian Y. Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: LLM as judge won't beat twice the data. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4264ee4376776907c0b87ed70b959585-Abstract-Conference.html.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021. doi: 10.1162/tacl_a_00373. URL https://aclanthology.org/2021.tacl-1.24/.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. FactKB: Generalizable factuality evaluation using language models enhanced with factual knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 933–952. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.59.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1396–1412, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.100.

Horacio Franco, Leonardo Neumeyer, Yoon Kim, and Orith Ronen. Automatic pronunciation scoring for language instruction. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pp. 1471–1474, 1997. doi: 10.1109/ICASSP.1997.596227.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman,

Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.wmt-1.2. URL https://aclanthology.org/2022.wmt-1.2/.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 578–628, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.51. URL https://aclanthology.org/2023.wmt-1.51/.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. Are LLMs breaking MT metrics? Results of the WMT24 metrics shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 47–81, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.2. URL https://aclanthology.org/2024.wmt-1.2/.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N. Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to evaluate reward models for RLHF. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2e01083b381b4865919b4915ef32e3d2-Abstract-Conference.html.

Mingqi Gao, Xinyu Hu, Li Lin, and Xiaojun Wan. Analyzing and evaluating correlation measures in NLG meta-evaluation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2199–2222, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.111. URL https://aclanthology.org/2025.naacl-long.111/.

Mingqi Gao, Anthony Sicilia, and Weiyan Shi. Which metrics save the most human annotation? Prediction-powered evaluation and meta-evaluation, 2026. URL https://arxiv.org/abs/2608.26638. Accepted at EMNLP 2026 (Main).

Sebastian Gehrmann. Reward models are metrics in a trench coat. arXiv:2510.03231, 2025.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Ariel Gera, Odellia Boni, Yotam Perlitz, Roy Bar-Haim, Lilach Eden, and Asaf Yehudai. JuStRank: Benchmarking LLM judges for system ranking. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 682–712, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.34. URL https://aclanthology.org/2025.acl-long.34/.

Leander Girrbach, Chi-Ping Su, Tankred Saanum, Richard Socher, Eric Schulz, and Zeynep Akata. Reference-free rating of LLM responses via latent information. arXiv:2509.24678, 2025.

Yuan Gong, Ziyi Chen, Iek-Heng Chu, Peng Chang, and James Glass. Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7262–7266, 2022. doi: 10.1109/ICASSP43922.2022.9746743. URL https://arxiv.org/abs/2205.03432.

Luke Guerdan, Solon Barocas, Kenneth Holstein, Hanna Wallach, Zhiwei Steven Wu, and Alexandra Chouldechova. Validating LLM-as-a-judge systems under rating indeterminacy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://neurips.cc/virtual/2025/poster/117308.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, 2018. doi: 10.18653/v1/N18-2017.

Ellen L. Hamaker. The curious case of the cross-sectional correlation. *Multivariate Behavioral Research*, 59(6):1111–1122, 2024. doi: 10.1080/00273171.2022.2155930. Published online 2023.

David J. Hand. Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1):103–123, 2009. doi: 10.1007/s10994-009-5119-5.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. arXiv:2103.03874.

L. Jeff Hong, Weiwei Fan, and Jun Luo. Review on ranking and selection: A new perspective. *Frontiers of Engineering Management*, 8(3):321–343, 2021. doi: 10.1007/s42524-021-0152-6.

Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. Themis: A reference-free NLG evaluation language model with flexibility and interpretability. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15924–15951. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.891.

Xinyu Hu, Mingqi Gao, Li Lin, Zhenghan Yu, and Xiaojun Wan. A dual-perspective NLG meta-evaluation framework with automatic benchmark and better interpretability. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27372–27395, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1327. URL https://aclanthology.org/2025.acl-long.1327/.

Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. The VoiceMOS challenge 2022. In *Proc. Interspeech 2022*, pp. 4536–4540, 2022a. doi: 10.21437/Interspeech.2022-970.

Wen-Chin Huang, Erica Cooper, Junichi Yamagishi, and Tomoki Toda. LDNet: Unified listener dependent modeling in MOS prediction for synthetic speech. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 896–900, 2022b. doi: 10.1109/ICASSP43922.2022.9747222.

ITU-T. Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. Recommendation ITU-T P.1401 (01/2020), International Telecommunication Union, Geneva, 2020. URL https://www.itu.int/rec/T-REC-P.1401-202001-I/en.

Justin Kay, Grant Van Horn, Subhransu Maji, Daniel Sheldon, and Sara Beery. Consensus-driven active model selection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4594–4604, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Kay_Consensus-Driven_Active_Model_Selection_ICCV_2025_paper.html.

Alex Kim, Keonwoo Kim, and Sangwon Yoon. DEBATE: Devil's advocate-based assessment and text evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1885–1897. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.112.

Sunghwan Kim, Dongjin Kang, Taeyoon Kwon, Hyungjoo Chae, Dongha Lee, and Jinyoung Yeo. Rethinking reward model evaluation through the lens of reward overoptimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13252–13280, 2025. doi: 10.18653/v1/2025.acl-long.649.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz (eds.), *Proceedings of the Sixth Conference on Machine Translation*, pp. 478–494, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.57/.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pp. 1440–1453, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.131.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Tenth Conference on Machine Translation*, pp. 355–413, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.wmt-1.22. URL https://aclanthology.org/2025.wmt-1.22/.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025. doi: 10.18653/v1/2025.findings-naacl.96.

Eddie Landesberg. When LLM judge scores look good but Best-of-N decisions fail, 2026. URL https://arxiv.org/abs/2603.12520. arXiv preprint arXiv:2603.12520.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Tenth Conference on Machine Translation*, pp. 436–483, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-341-8. doi: 10.18653/v1/2025.wmt-1.24. URL https://aclanthology.org/2025.wmt-1.24/.

Yichong Leng, Xu Tan, Sheng Zhao, Frank Soong, Xiang-Yang Li, and Tao Qin. MBNet: MOS prediction for synthesized speech with mean-bias network. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 391–395, 2021. doi: 10.1109/ICASSP39728.2021.9413877.

Yanran Li. Calibrate, don't curate: Label-efficient estimation from noisy LLM judges, 2026. URL https://arxiv.org/abs/2605.09702.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2305.20050.

Zheng Wei Lim, Ekaterina Vylomova, Trevor Cohn, and Charles Kemp. Simpson's paradox and the accuracy-fluency tradeoff in translation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 92–103, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.9. URL https://aclanthology.org/2024.acl-short.9/.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? Rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21558–21572, 2023a. doi: 10.52202/075280-0943.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. HD-Eval: Aligning large language model evaluators through hierarchical criteria decomposition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7641–7660. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.413.

Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. MOSNet: Deep learning-based objective assessment for voice conversion. In *Proc. Interspeech 2019*, pp. 1541–1545, 2019. doi: 10.21437/Interspeech.2019-2003.

Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. In *Proc. Odyssey 2018*, pp. 195–202, 2018. URL https://arxiv.org/abs/1804.04262.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. RewardBench 2: Advancing reward model evaluation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://iclr.cc/virtual/2026/poster/10008221.

Georgia Maniati, Alexandra Vioni, Nikolaos Ellinas, Karolos Nikitaras, Konstantinos Klapsas, June Sig Sung, Gunu Jho, Aimilios Chalamandaris, and Pirros Tsiakoulis. SOMOS: The Samsung open MOS dataset for the evaluation of neural text-to-speech synthesis. In *Proc. Interspeech 2022*, pp. 2388–2392, 2022. doi: 10.21437/Interspeech.2022-10922.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4984–4997, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.448. URL https://aclanthology.org/2020.acl-main.448/.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations, 2024. URL https://arxiv.org/abs/2411.00640.

Christoph Minixhofer, Ondrej Klejch, and Peter Bell. TTSDS2: Resources and benchmark for evaluating human-quality text to speech systems. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=uGai5lYHlV.

Yair Mundlak. On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85, 1978. doi: 10.2307/1913646.

Leonardo Neumeyer, Horacio Franco, Mitchel Weintraub, and Patti Price. Automatic text-independent pronunciation scoring of foreign language student speech. In *Proc. 4th International Conference on Spoken Language Processing (ICSLP 96)*, pp. 1457–1460, 1996. doi: 10.21437/ICSLP.1996-372.

Justin D. Norman, Michael U. Rivera, and D. Alex Hughes. Reliability without validity: A systematic, large-scale evaluation of LLM-as-a-judge models across agreement, consistency, and bias, 2026. URL https://arxiv.org/abs/2606.19544.

Patrik Okanovic, Andreas Kirsch, Jannes Kasper, Torsten Hoefler, Andreas Krause, and Nezihe Merve Gürel. All models are wrong, some are useful: Model selection with limited labels. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2035–2043, 2025. URL https://proceedings.mlr.press/v258/okanovic25a.html.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4812–4829, 2021. doi: 10.18653/v1/2021.naacl-main.383. URL https://aclanthology.org/2021.naacl-main.383/.

Karl Pearson. *Mathematical Contributions to the Theory of Evolution. XIV. On the General Theory of Skew Correlation and Non-Linear Regression*. Number II in Drapers' Company Research Memoirs, Biometric Series. Dulau and Co., London, 1905.

Stefano Perrella, Lorenzo Proietti, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. Beyond correlation: Interpretable evaluation of machine translation metrics. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20689–20714, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1152. URL https://aclanthology.org/2024.emnlp-main.1152/.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16216–16244, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.856. URL https://aclanthology.org/2024.acl-long.856/.

Jaden Pieper and Stephen D. Voran. Bounds on agreement between subjective and objective measurements. *IEEE Transactions on Multimedia*, 2026. doi: 10.1109/TMM.2026.3712414. Early access.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, pp. 180–191, 2018. doi: 10.18653/v1/S18-2023.

Lorenzo Proietti, Stefano Perrella, and Roberto Navigli. Has machine translation evaluation achieved human parity? The human reference and the limits of progress. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 790–813, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.63. URL https://aclanthology.org/2025.acl-short.63/.

Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason D. Lee, and Sanjeev Arora. What makes a reward model a good teacher? An optimization perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://neurips.cc/virtual/2025/poster/119671.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A

benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98. arXiv:2311.12022.

W. S. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357, 1950. doi: 10.2307/2087176.

Jon Saad-Falcon, E. Kelly Buchanan, Mayee F. Chen, Tzu-Heng Huang, Brendan McLaughlin, Tanvir Bhathal, Shang Zhu, Ben Athiwaratkun, Frederic Sala, Scott Linderman, Azalia Mirhoseini, and Christopher Ré. Weaver: Shrinking the generation-verification gap by scaling compute for verification. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2639. arXiv:2506.18203.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *Proc. Interspeech 2022*, pp. 4521–4525, 2022. doi: 10.21437/Interspeech.2022-439.

David Salinas, Omar Swelam, and Frank Hutter. Tuning LLM judge design decisions for 1/1000 of the cost. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 52728–52744, 2025. URL https://proceedings.mlr.press/v267/salinas25a.html.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

C. Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3(3):271–295, 1910. doi: 10.1111/j.2044-8295.1910.tb00206.x.

Julius Steen and Katja Markert. How to find strong summary coherence measures? A toolbox and a comparative study for summary coherence measure evaluation. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 6035–6049, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.527/.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM2)*, pp. 404–430, 2025. URL https://aclanthology.org/2025.gem-1.33/.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 1222–1234, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.118. URL https://aclanthology.org/2024.wmt-1.118/.

Wei-Cheng Tseng, Chien-yu Huang, Wei-Tsung Kao, Yist Y. Lin, and Hung-yi Lee. Utilizing self-supervised representations for MOS prediction. In *Proc. Interspeech 2021*, pp. 2781–2785, 2021. doi: 10.21437/Interspeech.2021-2013.

Shengxin Tu, Chun Li, and Bryan E. Shepherd. Between- and within-cluster Spearman rank correlations. *Statistics in Medicine*, 44(3-4):e10326, 2025. doi: 10.1002/sim.10326. URL https://arxiv.org/abs/2402.11341.

Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. arXiv:2404.18796, 2024.

Alexandra Vioni, Georgia Maniati, Nikolaos Ellinas, June Sig Sung, Inchul Hwang, Aimilios Chalamandaris, and Pirros Tsiakoulis. Investigating content-aware neural text-to-speech MOS prediction using prosodic and linguistic features. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096255.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. CAM++: A fast and efficient network for speaker verification using context-aware masking. In *Proc. Interspeech 2023*, pp. 5301–5305, 2023a. doi: 10.21437/Interspeech.2023-1513.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is ChatGPT a good NLG evaluator? A preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pp. 1–11, 2023b. doi: 10.18653/v1/2023.newsum-1.1. URL https://arxiv.org/abs/2303.04048.

Kexin Wang, Yunlong Zhao, Qianqian Dong, Tom Ko, and Mingxuan Wang. MOSPC: MOS prediction based on pairwise comparison. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1547–1556, 2023c. doi: 10.18653/v1/2023.acl-short.132.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, Bangkok, Thailand, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Victor Wang, Michael J. Q. Zhang, and Eunsol Choi. Improving LLM-as-a-judge inference with the judgment distribution. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23173–23199, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1259. URL https://aclanthology.org/2025.findings-emnlp.1259/. arXiv:2503.03064.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 95266–95290, 2024b. doi: 10.52202/079017-3018.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer 2: Open-source dataset for training top-performing reward models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 1474–1501, 2024c. URL https://papers.nips.cc/paper_files/paper/2024/hash/02fd91a387a6a5a5751e81b58a75af90-Abstract-Datasets_and_Benchmarks_Track.html.

Xueru Wen, Jie Lou, Yaojie Lu, Hongyu Lin, Xing Yu, Xinyu Lu, Ben He, Xianpei Han, Debing Zhang, and Le Sun. Rethinking reward model evaluation: Are we barking up the wrong tree? In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/93f250215e4889119807b6fac3a57aec-Abstract-Conference.html.

Tianruo Rose Xu, Vedant Gaur, Liu Leqi, and Tanya Goyal. The progress illusion: Revisiting meta-evaluation standards of LLM evaluators. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19033–19043, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1036. URL https://aclanthology.org/2025.findings-emnlp.1036/.

Yang Xu, Jiefu Zhang, Haixiang Sun, Zihan Zhou, Tianyu Cao, and Vaneet Aggarwal. Towards reliable LLM evaluation: Correcting the winner's curse in adaptive benchmarking, 2026. URL https://arxiv.org/abs/2605.05973.

Bi-Cheng Yan and Berlin Chen. An effective hierarchical graph attention network modeling approach for pronunciation assessment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3974–3985, 2024. doi: 10.1109/TASLP.2024.3449111.

Bi-Cheng Yan, Ming-Kang Tsai, and Berlin Chen. MuFFIN: Multifaceted pronunciation feedback model with interactive hierarchical neural modeling. *IEEE Transactions on Audio, Speech and Language Processing*, 33:4295–4310, 2025a. doi: 10.1109/TASLPRO.2025.3619765.

Bi-Cheng Yan, Hsin Wei Wang, Fu-An Chao, Tien-Hong Lo, Yung-Chang Hsu, and Berlin Chen. HiPPO: Exploring a novel hierarchical pronunciation assessment approach for spoken languages. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 810–823, Mumbai, India, 2025b. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. doi: 10.18653/v1/2025.ijcnlp-long.45. URL https://aclanthology.org/2025.ijcnlp-long.45/.

Bi-Cheng Yan, Yi-Cheng Wang, Jiun-Ting Li, Meng-Shin Lin, Hsin-Wei Wang, Wei-Cheng Chao, and Berlin Chen. ConPCO: Preserving phoneme characteristics for automatic pronunciation assessment leveraging contrastive ordinal regularization. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025c. doi: 10.1109/ICASSP49660.2025.10890778.

Takenori Yoshimura, Gustav Eje Henter, Oliver Watts, Mirjam Wester, Junichi Yamagishi, and Keiichi Tokuda. A hierarchical predictor of synthetic speech naturalness using neural networks. In *Proc. Interspeech 2016*, pp. 342–346, 2016. doi: 10.21437/Interspeech.2016-847.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. BARTScore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27263–27277, 2021. URL https://arxiv.org/abs/2106.11520.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. speechocean762: An open-source non-native English speech corpus for pronunciation assessment. In *Proc. Interspeech 2021*, pp. 3710–3714, 2021. doi: 10.21437/Interspeech.2021-1259.

Shiyue Zhang and Mohit Bansal. Finding a balanced degree of automation for summary evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6617–6632. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.531.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10495–10516, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.547. URL https://aclanthology.org/2025.findings-acl.547/.

Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Jing Jiang, and Min Lin. Cheating automatic LLM benchmarks: Null models achieve high win rates. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9adc8ada9183f4b9a007a02773fd8114-Abstract-Conference.html.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2023–2038, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.131. URL https://aclanthology.org/2022.emnlp-main.131/.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911. arXiv preprint arXiv:2311.07911.

Yilun Zhou, Austin Xu, Peifeng Wang, Caiming Xiong, and Shafiq Joty. Evaluating judges as evaluators: The JETTS benchmark of LLM-as-judges as test-time scaling evaluators. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79436–79471, 2025. URL https://proceedings.mlr.press/v267/zhou25af.html.

Bin Zhu, Yi Xie, and Yanghui Rao. A finite-calibration regime map for LLM judge panels, 2026. URL https://arxiv.org/abs/2606.01034. Accepted at WISE 2026.

Maike Züfle, Vilém Zouhar, Tu Anh Dinh, Felipe Maia Polo, Jan Niehues, and Mrinmaya Sachan. COMET-poly: Machine translation metric grounded in other candidates. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Tenth Conference on Machine Translation*, pp. 887–904, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-341-8. doi: 10.18653/v1/2025.wmt-1.63. URL https://aclanthology.org/2025.wmt-1.63/.

A TERMS

Cluster, candidate. The items among which one decision is taken: a prompt's responses, a source segment's translations, a document's summaries, a speaker's utterances. A candidate pair is two items of one cluster.

Evaluator, panel. A metric, reward model, LLM judge or scorer; the panel is the evaluators a benchmark scores on every item of a configuration, and on WMT the pool is the systems whose outputs are kept.

Configuration, cell, set. A configuration is one benchmark, label and panel; a cell is a configuration at one label budget and payoff; a set is the configurations a result averages with equal weight: the exploratory set of Section 4, the validation sets of the first protocol, and the frozen sets of the later protocols.

Payoff. Best of n, the share of the achievable label gain recovered by the candidate an evaluator ranks first, and the pairwise choice, the share of decisive candidate pairs it orders as the labels do (Appendix H).

Rule. What selects one evaluator from the labelled clusters: pooled Pearson r_T, within-cluster Pearson r_W, the grouped mean grp of per-cluster Spearman coefficients; the pairwise accuracy and best-of-n recovery inside the labelled clusters, each of which is dec, the payoff's own decision statistic, under its own payoff; and own, the benchmark's headline statistic (Appendix G).

Budget. The labelled clusters a rule reads: a number on the grid $\{2, 3, 4, 6, 8, 12, 16, 24, 32, 48, \dots\}$ of Appendix K up to half a configuration's clusters, or half of them, a few dozen to nearly a thousand by configuration; eight and half are reported most.

Advantage, gain, regret. The advantage is the points of pooling's selection error over evaluator pairs a rule removes; the gain, the points of payoff its pick earns over pooling's on the evaluation half; a rule's regret, the points by which its pick falls short of the evaluation half's best evaluator. Intervals of paired utility differences are the estimate plus or minus 1.96 bootstrap standard deviations unless stated otherwise; all are descriptive.

B PROOFS

Equation 1. Write $\tilde{p} = p - \bar{p} = p^B + p^W$ and $\tilde{y} = y^B + y^W$. For every cluster the within parts sum to zero, and the between parts are constant within a cluster, so $\langle p^B, y^W \rangle = \sum_c (\bar{p}_c - \bar{p}) \sum_{i \in c} y_i^W = 0$, and likewise $\langle p^W, y^B \rangle = 0$. Hence $\langle \tilde{p}, \tilde{y} \rangle = \langle p^B, y^B \rangle + \langle p^W, y^W \rangle$ and $\|\tilde{v}\|^2 = \|v^B\|^2 + \|v^W\|^2$. Dividing by $\|\tilde{p}\| \, \|\tilde{y}\|$ gives

$$r_T = \frac{\|p^B\| \|y^B\|}{\|\tilde{p}\| \|\tilde{y}\|} \, r_B + \frac{\|p^W\| \|y^W\|}{\|\tilde{p}\| \|\tilde{y}\|} \, r_W = \eta_p \eta_y \, r_B + \sqrt{1 - \eta_p^2} \sqrt{1 - \eta_y^2} \, r_W .$$

No assumption on cluster sizes is used.

Corollary 1. *Let both vectors have positive variance and* $0 < \eta_p, \eta_y < 1$*, so that* $D = \sqrt{1-\eta_p^2}\sqrt{1-\eta_y^2} > 0$*. Given* η_p *and* η_y*, a pooled correlation* r_T *restricts* r_W *to* $[(r_T - \eta_p\eta_y)/D,\ (r_T + \eta_p\eta_y)/D] \cap [-1, 1]$*, and to nothing narrower whenever* $C \geq 3$ *and* $N - C \geq 2$*; the interval is all of* $[-1, 1]$ *exactly when* $|r_T| \leq \eta_p\eta_y - D$*.*

Proof of Corollary 1. Equation 1 is linear in r_B, so with η_p, η_y and r_T fixed the admissible r_W are the $w \in [-1, 1]$ for which some $b \in [-1, 1]$ gives $r_T = \eta_p\eta_y b + Dw$, which is the stated interval. Every point of it is attained, so it cannot be narrowed, whenever b and w can be chosen independently over $[-1, 1]$: the between-cluster and within-cluster subspaces are orthogonal, and a correlation inside a subspace of dimension at least two can be set to any value in $[-1, 1]$. Those dimensions are $C - 1$ and $N - C$, which is where $C \geq 3$ and $N - C \geq 2$ come from. The condition is sufficient for sharpness at every feasible triple and is not necessary at a given one: a one-dimensional subspace forces its correlation to ± 1, which leaves the interval loose in general, but at a boundary point it can still be exact. A witness: two clusters of two items, $b = (-1, -1, 1, 1)/2$, $w = (-1, 1, -1, 1)/2$, $y = 0.8b + 0.6w$ and $p = 0.5b + \sqrt{0.75}\, w$ give $\eta_y^2 = 0.64$, $\eta_p^2 = 0.25$ and $r_T = 0.920$, where the interval is the single point $\{1\}$ and $r_W = 1$ (`scripts/sim/boundary_checks.py`). If $\eta_p = 0$ then $p^B = 0$, r_B is a cosine with the zero vector and is undefined, the first term of Equation 1 vanishes, and $r_W = r_T/D$ exactly, which is the interval's own limit; $\eta_y = 0$ is the same statement with the roles exchanged, and `scripts/sim/identification_witness.py` checks it to $1e-17$. The interval covers all of $[-1, 1]$ exactly when $(r_T - \eta_p\eta_y)/D \leq -1$ and $(r_T + \eta_p\eta_y)/D \geq 1$, that is when $|r_T| \leq \eta_p\eta_y - D$; this needs $\eta_p\eta_y \geq D$, and squaring gives $\eta_p^2\eta_y^2 \geq (1-\eta_p^2)(1-\eta_y^2)$, which reduces to $\eta_p^2 + \eta_y^2 \geq 1$ (checked on a 9801-point grid with no exception). The condition is therefore on the two shares together, and a between-cluster label share above one half is what it requires only of an evaluator whose own share equals the label's. The label does not decide alone: at $\eta_y^2 = 0.25$, $\eta_p^2 = 0.99$ and $r_T = 0.40$ the slack $\eta_p\eta_y - D - |r_T|$ is 0.011, and the same script builds two data sets carrying exactly those three numbers, one with $r_W = +1$ at $r_B = 0.630$ and one with $r_W = -1$ at $r_B = 0.978$, both to $2e-16$. □

Item-blind predictors and the ceiling. Call a predictor *item-blind* if p_i depends on the cluster and on its other items but not on item i; it may use their labels, and the oracles of this paper do, so no value it reaches says what cluster identity by itself supplies. A *cluster-constant* predictor takes one value per cluster. The bound below is Pearson's correlation ratio of the labels on cluster identity (Pearson, 1905), computed before any evaluator is trained, and it is in sample: with many clusters of n items the in-sample share tends to $\rho + (1-\rho)/n$ for a population intraclass correlation ρ, so both the in-sample η_y and the population ceiling $\sqrt{\rho}$ are reported (Appendix D); rater noise bounds r_W by $\sqrt{\rho_W}$, the within-cluster reliability.

Proposition 1. *For a cluster-constant predictor of positive variance and a label with* $\eta_y > 0$*,* $r_T = \eta_y\, r_B \leq \eta_y$*, with equality when* p *is an increasing affine function of the cluster means of* y*; if* $\eta_y = 0$ *then* r_B *is undefined and* $r_T = 0$*.*

Proof. If p is constant within clusters then $p^W = 0$ and $\eta_p = 1$, so $r_T = \eta_y r_B$ and $|r_B| \leq 1$. The bound is attained if and only if $p^B = a\, y^B$ with $a > 0$, that is $p_i = \alpha + a\, \bar{y}_{c(i)}$. □

In-sample share and intraclass correlation. Under $y_i = \mu + u_{c(i)} + e_i$ with independent $u_c \sim (0, \sigma_B^2)$ and $e_i \sim (0, \sigma_W^2)$, the between and within sums of squares satisfy $\mathbb{E}[\mathrm{SSB}] = (C-1)(n_0\sigma_B^2 + \sigma_W^2)$ and $\mathbb{E}[\mathrm{SSW}] = (N-C)\sigma_W^2$, with $n_0 = (N - \sum_c n_c^2/N)/(C-1)$. For balanced clusters of size n and many clusters, $\eta_y^2 = \mathrm{SSB}/(\mathrm{SSB} + \mathrm{SSW}) \to \rho + (1-\rho)/n$. The estimator $\hat{\rho} = (\mathrm{MSB} - \mathrm{MSW})/(\mathrm{MSB} + (n_0 - 1)\mathrm{MSW})$ removes the $(1-\rho)/n$ term. The limit is a large-C statement and not a finite-sample expectation: with Gaussian labels and $\rho = 0$, SSB/SST is a $\mathrm{Beta}\big((C-1)/2,\, C(n-1)/2\big)$ variable with mean $(C-1)/(Cn-1)$, which at $n = 2$ is $(C-1)/(2C-1) < \frac{1}{2}$ for every C, 0.400 at three clusters and 0.498 at 125; with $\rho > 0$ the finite-sample mean also sits below the limit $(1+\rho)/2$, at 0.71 for $\rho = 0.5$ and nine clusters against 0.75 (`scripts/sim/boundary_checks.py`). Nothing stronger than the limit holds: the share at $n = 2$ does not exceed a half in every sample, which three clusters labelled $(-1, 1)$ show, nor in expectation whatever ρ is, which the Beta mean shows.

Proposition 2. *Under the random-effects model, an evaluator that outputs the mean label of k other items of the same cluster has correlation $\rho/\sqrt{\rho + (1-\rho)/k}$ with the label. This is $\sqrt{\rho\,\rho_k}$ with ρ_k the Spearman–Brown reliability of a mean of k items (Spearman, 1910; Brown, 1910).*

Proof. Let $p_i = k^{-1}\sum_{j\in H_i} y_j$ with H_i a set of k items of cluster $c(i)$ that excludes i. Then $\mathrm{Cov}(p_i, y_i) = \mathrm{Var}(u_c) = \sigma_B^2$, $\mathrm{Var}(p_i) = \sigma_B^2 + \sigma_W^2/k$ and $\mathrm{Var}(y_i) = \sigma_B^2 + \sigma_W^2$, so

$$\mathrm{corr}(p_i, y_i) = \frac{\sigma_B^2}{\sqrt{(\sigma_B^2 + \sigma_W^2/k)(\sigma_B^2 + \sigma_W^2)}} = \frac{\rho}{\sqrt{\rho + (1-\rho)/k}}.$$

The sample pooled correlation converges to this value as the number of clusters grows. Appendix D checks the approximation at the sizes used in this paper. The same expression is $\sqrt{\rho\,\rho_k}$ with $\rho_k = k\rho/(1 + (k-1)\rho)$ the Spearman–Brown reliability of a mean of k items (Spearman, 1910; Brown, 1910), so the profile curve is the classical attenuation bound (Spearman, 1904) for a predictor whose true score is the cluster mean. □

Proposition 3. *Let $p_i = (S_{c(i)} - y_i)/(n_{c(i)} - 1)$, S_c the sum of the labels of cluster c, and let y vary between and within clusters. Then $\bar{p}_c = \bar{y}_c$ for every cluster, and if all clusters hold n items, $r_B = 1$, $r_W = -1$ and $r_T = \big(\eta_y^2 - (1-\eta_y^2)/(n-1)\big)\big/\sqrt{\eta_y^2 + (1-\eta_y^2)/(n-1)^2}$.*

Proposition 3. With $p_i = (S_c - y_i)/(n_c - 1)$ for $i \in c$, the cluster mean of p is $\sum_{i\in c} p_i/n_c = (n_c S_c - S_c)/(n_c(n_c-1)) = \bar{y}_c$, so $p^B = y^B$. Subtracting, $p_i - \bar{p}_c = (S_c - n_c y_i)/(n_c(n_c-1)) = -(y_i - \bar{y}_c)/(n_c - 1)$, so $p^W = -y^W/(n-1)$ when every cluster holds n items. Hence $r_B = 1$, $r_W = -1$, $\langle \tilde{p}, \tilde{y}\rangle = \|y^B\|^2 - \|y^W\|^2/(n-1)$ and $\|\tilde{p}\|^2 = \|y^B\|^2 + \|y^W\|^2/(n-1)^2$. Dividing by $\|\tilde{p}\|\|\tilde{y}\|$ and writing $\|y^B\|^2/\|\tilde{y}\|^2 = \eta_y^2$ gives Proposition 3. For unbalanced clusters the same two displays hold cluster by cluster with n_c in place of n, so $p^B = y^B$ and $r_B = 1$ still hold, but r_T is then a ratio of sums over clusters and is no longer a function of η_y and the sizes alone: with ten clusters of two items and ten of twenty and η_y^2 held at 0.600, moving the within-cluster variation between the small and the large clusters moves r_T between 0.208 and 0.732, with r_W at -0.991 where r_T is smallest and -0.475 where it is largest (`scripts/sim/prop_limits.py`). Proposition 3 is therefore stated for equal cluster sizes and computed directly otherwise. The negative within-cluster part belongs to this predictor and not to the class it comes from: negating it is item-blind under the same definition and has $r_W = +1$, because the within-cluster deviations of a finite cluster sum to zero and the other items therefore locate the scored one. Item-blindness does not prevent a predictor from ordering items inside a cluster; what the leave-one-out mean does is order them backwards, and Equation 1 charges it for that.

The leave-one-out mean does not maximise over the item-blind class. Proposition 3 gives a constructible value, not a supremum, and the distinction matters because the predictor's within-cluster correlation is fixed at -1: it inverts the within-cluster ordering rather than ignoring it, and pays the within-cluster term of Equation 1 for doing so. Take three clusters of four items with labels $(0,0,0,3)$, $(3,3,3,6)$ and $(6,6,6,9)$. The leave-one-out mean reaches a pooled correlation of 0.788. The leave-one-out *median* is item-blind under the same definition, since each item is scored from the other three of its cluster, but the ties make it constant on each cluster, so it is also cluster-constant and reaches $\eta_y = 0.883$, the ceiling of Proposition 1 and the value of the full cluster mean. No claim of optimality over item-blind predictors is therefore made anywhere in this paper: Proposition 1 bounds the cluster-constant subclass and Proposition 3 instantiates one predictor a benchmark holder can always compute without fitting anything.

Crossed designs. Let the centred label space be the orthogonal sum of subspaces $V_1, \ldots, V_m$ with projectors P_k, for example the row main effects, column main effects and interaction of a complete two-way layout. The argument for Equation 1 gives $r_T = \sum_k \sqrt{s_k(p)\,s_k(y)}\,\cos(P_k p, P_k y)$ with $s_k(v) = \|P_k\tilde{v}\|^2/\|\tilde{v}\|^2$. A predictor that depends on factor j only lies in V_j after centring and satisfies $r_T \le \sqrt{s_j(y)}$. Correlation computed over all cells after removing the row effects, which is the pooled row-centred coefficient, is bounded for a column-identity predictor by the square root of the column share of that complement.

This bound does not transfer to the per-row protocols, which have a different one. Per-document and per-segment correlation averages one correlation per row, and averaging discards the scale of

each row, which the pooled coefficient weights by. A rank-one label matrix $y_{rc} = \lambda_r a_c$ makes every row's correlation with the column-identity predictor a equal to one whatever the shares are: with $\lambda_r > 0$ the mean per-row correlation is 1.000 at three values of the spread of λ while the column-share bound is 0.998, 0.986 and 0.947, and the pooled row-centred coefficient equals that bound to four decimals in each case, which locates the failure exactly in the swap of pooling for averaging (`scripts/sim/prop_limits.py`). It does not follow that no label-only bound exists under a per-row protocol: the construction shows only that no bound below one holds for *every* label matrix, whereas a benchmark's reader holds one particular matrix. The maximum for that matrix is the following.

Proposition 4. *Fix a complete label matrix and a per-row protocol that averages one coefficient over the rows whose labels are not constant. Let u_r be the centred and normalised label vector of row r, or of its midranks under Spearman, and let m be the mean of u_r over those rows. Under a rank coefficient a cluster-constant predictor enters the average only through the weak order it induces on the clusters, while under Pearson it enters through its own values; the maximum of the mean per-row coefficient over all such predictors is*

- *$\|m\|$ for Pearson, attained at $p \propto m$ when $m \neq 0$; when $m = 0$ the maximum is zero, $p \propto m$ is not a prediction vector, and every non-constant predictor attains it;*
- *the maximum of $\langle \tilde{a}, m \rangle$ over ordered partitions of the clusters for Spearman, where a is the midrank vector of the partition and $\tilde{a}$ its centred and normalised form. The blocks of an optimal partition are consecutive in the order of m and the squared norm of the centred a depends on the block sizes alone, so for each value of one scalar parameter the maximum is a dynamic program over cut positions in the sorted m, quadratic in C; the parameter is searched by the hull enumeration below, which terminates;*
- *the maximum of $\sum_{i<j} \operatorname{sgn}(p_i - p_j) w_{ij} / \sqrt{n_0 - t_p}$ over ordered partitions for Kendall's τ_b, with $n_0 = \binom{C}{2}$, $w_{ij} = \frac{1}{R} \sum_r \operatorname{sgn}(y_{ri} - y_{rj}) / \sqrt{n_0 - t_r}$, t_r the tied label pairs of row r and t_p the tied pairs of the predictor. A dynamic program over subsets solves it exactly at the cluster counts here.*

Proof. A cluster-constant predictor supplies one vector p over the C clusters and every row sees the same p. Under Pearson, the row's coefficient is $\langle \hat{p}, u_r \rangle$ with $\hat{p}$ the centred and normalised p, so the mean is $\langle \hat{p}, \frac{1}{R} \sum_r u_r \rangle = \langle \hat{p}, m \rangle$, which Cauchy–Schwarz bounds by $\|m\|$ with equality at $\hat{p} = m/\|m\|$; m is centred, so it is an admissible predictor. Under Spearman the predictor enters through its midrank vector a, which is determined by the weak order it induces and by nothing else, and the mean is $\langle \tilde{a}, m \rangle$. If two clusters in different blocks are ordered against m, exchanging them leaves every block size and so the norm of $\tilde{a}$, and raises the numerator by the product of the two positive differences, so an optimal partition is consecutive in the order of m: what remains is a choice of cut positions. Both the numerator and the squared norm are sums of terms each depending on one block alone, so for a fixed λ the partition maximising numerator minus λ times squared norm is a shortest-path problem. A maximiser of $N/\sqrt{D}$ also maximises $N - \lambda D$ at $\lambda = N/(2D)$, since for any competitor $N' \leq (N/\sqrt{D})\sqrt{D'}$ gives $N' - \lambda D' \leq \frac{N}{\sqrt{D}} \sqrt{D}(t - t^2/2) \leq N - \lambda D$ with $t = \sqrt{D'/D}$, so every maximiser is a vertex of the upper convex hull of the attainable (D, N) and bisection on λ enumerates them. Under τ_b the coefficient of row r is $\sum_{i<j} \operatorname{sgn}(p_i - p_j) \operatorname{sgn}(y_{ri} - y_{rj}) / \sqrt{(n_0 - t_p)(n_0 - t_r)}$, so the mean is the displayed ratio, whose numerator sums over the pairs the predictor separates and whose denominator counts the pairs it ties; both are sums over blocks, and the same hull argument applies with the subset program in place of the path program. □

Ties are the reason the optimisation is over weak orders and not over permutations of $1, \ldots, C$. A predictor that ties two clusters gives up the pairs it ties in the numerator and is given those same pairs back in the denominator, and the trade can pay. The smallest witness is the matrix whose every row is $(0, 0, 1)$: the tied predictor $(0, 0, 1)$ correlates 1 with every row, while the best strict order reaches only 0.866 under Spearman and 0.816 under τ_b. A verification that draws predictors from a continuous distribution cannot see this, since such a predictor is untied with probability one. On the 11 rank-protocol configurations of Table 72 allowing ties raises the maximum over the strict-order optimum by between 0.0004 and 0.0767, and the cluster mean falls up to 0.0805 below it.

Proposition 5. *Fix labels y and a clustering, and let the protocol correlate over all items in Spearman's ρ or Kendall's τ_b. A cluster-constant predictor enters only through the weak order it induces on the clusters, so its coefficient depends on that order and on the cluster sizes alone. Under Spearman the coefficient of any such predictor is at most $\eta_{\mathrm{rank}(y)}$, the correlation ratio of the*

label midranks on cluster identity; under τ_b it is at most $\big(\sum_{a<b} q_{ab}^2 n_a n_b\big)^{1/2}/(n_0 - t_y)^{1/2}$, with $w_{ab} = \sum_{i\in a, j\in b} \operatorname{sgn}(y_j - y_i)$, $q_{ab} = |w_{ab}|/(n_a n_b)$, $n_0 = \binom{N}{2}$ and t_y the tied label pairs. Neither bound need be attained, and the cluster-mean predictor need not attain the maximum either.

Proof. Spearman's ρ of p against y is the Pearson correlation of their midrank vectors, and the midrank vector of a cluster-constant p is cluster-constant, so Proposition 1 applied in midrank space gives the first bound; it drops the constraint that a midrank vector takes the values midranks take, which is why it need not be attained. For τ_b, an ordered partition of the clusters gives numerator $\sum_{\text{separated } a<b} \pm w_{ab} \le \sum_{\text{separated}} q_{ab} n_a n_b$ and untied pair count $X = \sum_{\text{separated}} n_a n_b$, and Cauchy–Schwarz gives $\sum q_{ab} n_a n_b \le (\sum q_{ab}^2 n_a n_b)^{1/2} X^{1/2}$, so the ratio to $\sqrt{X}$ is at most the root of the sum over *all* pairs. □

The three programs, stated apart. Three optimisations are solved here, each stated with its guarantee and its cost. *Per-row Spearman* (Proposition 4): the exchange argument above makes an optimal partition consecutive in the order of m, so for a fixed λ the partition maximising $N - \lambda D$ is a shortest path over cut positions at $O(C^2)$, and the result is exact; it runs at $C = 16$ and $C = 24$ here. *Per-row Kendall* (Proposition 4): the cross-block weights w_{ij} are not a function of one scalar per cluster, so no exchange argument orders the blocks, and for a fixed λ the program over ordered partitions is a dynamic program over subsets that places one block at a time, exact, at $O(3^C)$ per solve; it runs at $C = 10$ here and the code refuses $C > 22$. *Pooled rank coefficients* (Proposition 5): the path program over partitions consecutive in the order of the cluster-mean midrank is exact for Spearman at equal cluster sizes by the same exchange argument; with unequal sizes an exchange alters the block sizes and the argument does not close, and for τ_b the exchange condition between adjacent blocks is a pairwise majority, which can cycle. What the path program attains there is therefore reported as an attained value with the certified bound of the proposition beside it, and `scripts/sim/pooled_bound.py` records that on 3445 instances at three to five clusters with unequal sizes and tied labels, enumerating every weak order, the path program equals the Spearman maximum in every case and the τ_b maximum in 3443 of them, at most 0.0222 short, with a pinned counterexample of five clusters of sizes $1, 3, 3, 3, 1$ where it reaches 0.5703 against 0.5752. Table 2 gives all three numbers for the 9 pooled rank rows of Table 72: the cluster mean falls 0.0005 to 0.0702 below what the path program attains, which in turn falls 0.0021 to 0.0623 below the certified bound. *The outer enumeration*, common to the three: each coefficient has the form $N/\sqrt{D}$ with N and D additive over blocks; a maximiser of $N/\sqrt{D}$ maximises $N - \lambda^* D$ at $\lambda^* = N^*/(2D^*)$, as shown above, and so is a vertex of the upper convex hull of the attainable (D, N). The hull is enumerated by solving the inner program at $\lambda = 0$ and at $\lambda_{\max} = N_{\max}/(2D_{\min})$, its two ends, then at the slope of the chord joining two vertices already found: a solution whose value meets the chord proves no vertex lies strictly between and closes that interval, and any other solution is a new vertex and splits it in two. Every solve closes an interval or finds a vertex, the hull has finitely many vertices, so the recursion terminates after V solves and the total cost is V times the inner program's. Two caps guard against a pathological instance, a recursion depth of 40 and 2000 solves; a cap that bound would leave a lower bound rather than a maximum, so `scripts/sim/hull_search_audit.py` records what the enumeration did on every row the paper prints: over those 20 rows a median of 32 solves and at most 907, depth 15, and neither cap bound. With the search capped at 400 solves both bound on the three source-clustered WMT rows, and on one of them the capped search returns a value 0.00045 below what the completed enumeration reaches; Table 2 carries the completed figure. The witness of the main text is reproduced in the same script.

All three forms are checked in `scripts/sim/per_row_bound.py`. On 60 random matrices per coefficient at three to five clusters every weak order is enumerated and none beats the claimed maximum; on 300 further matrices per coefficient the claimed maximiser attains the claimed value to machine precision, and 450 random predictors per matrix, a third untied and two thirds drawn from two- and three-level alphabets, never beat it; the grouped coefficient itself agrees with `scipy` to machine precision on tied predictors, so the check does not rest on our own implementation of the coefficient; and the two-document instance $y_1 = (0, 1, 2)$, $y_2 = (2, 1, 0)$, where the maximum is exactly zero, the rank-one matrix, where it is one, and the tied witness above are reproduced.

Table 2: The rows of Table 72 whose protocol is a rank coefficient pooled over items, where the bound is certified rather than attained. Cluster mean is the coefficient of the predictor that outputs each cluster's mean label, which Proposition 1 makes the maximum under Pearson and which is one admissible predictor here. Attained is the best value the path program of Appendix B finds over ordered partitions of the clusters, so some label-only predictor really reaches it. Bound is the certified upper bound of Proposition 5 and is what Table 72 prints.

Benchmark	Label	Cluster	Protocol	C	Cluster mean	Attained	Bound
BVCC	mos	system	Spearman	187	0.873	0.873	0.876
SOMOS	mos	system	Spearman	201	0.571	0.573	0.577
VCC2018	mos	system	Spearman	26	0.520	0.523	0.564
WMT22 en-de	mqm	system	τ_b	13	0.083	0.095	0.105
WMT22 en-de	mqm, source clusters	source segment	τ_b	1315	0.577	0.647	0.709
WMT22 zh-en	mqm	system	τ_b	14	0.096	0.138	0.151
WMT22 zh-en	mqm, source clusters	source segment	τ_b	1875	0.561	0.621	0.680
WMT22 en-ru	mqm	system	τ_b	15	0.155	0.171	0.189
WMT22 en-ru	mqm, source clusters	source segment	τ_b	1315	0.580	0.643	0.702

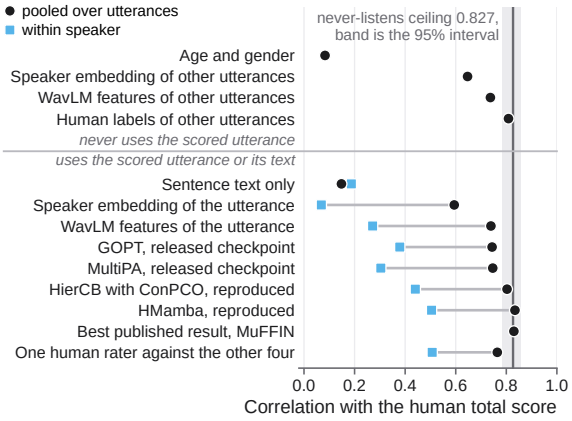


Figure 3: **Predictors of the speechocean762 total score.** Pearson correlation with the released total score on the speechocean762 test set. Predictors above the line use only other utterances of the same speaker: fit on training speakers, shown a random half of each test speaker's utterances and scored on the other half over 20 partitions. Their prediction is constant inside a speaker, so the within-speaker term of Equation 1 is zero while the within-speaker correlation itself is undefined, which is how the toolkit reports it; a cluster-constant predictor contributes nothing to the within-cluster component rather than contributing a correlation of zero. Predictors below it use the scored utterance or its sentence text and are scored on all test utterances. The vertical line is the population ceiling; the human row correlates one rater with the unrounded median of the other four, which is one rater fewer and one rounding step less than the released label.

C SPEECHOCEAN762 IN DETAIL

Table 68 gives the label-only statistics of all four labels. Proposition 2 predicts the pooled correlation of a speaker-mean predictor built from k other labelled utterances within 0.012 for every aspect and every k from 1 to 10, and the four reproduced systems land within 0.009 of their published pooled correlations apart from the released MultiPA checkpoint at 0.015 (Table 70). Age and gender alone reach 0.083 and a speaker-verification embedding (Wang et al., 2023a) of the speaker's other utterances 0.647. The decomposition separates what listening evaluators gain from the speaker and from the utterance. The ridge regression on the scored utterance has a between-speaker correlation of 0.893 against a within-speaker correlation of 0.271, and a speaker-verification embedding of the scored utterance reaches 0.594 pooled with 0.069 within. Surface features of the sentence text reach 0.187 within speakers without any audio, and partialling out their prediction lowers the WavLM regression to 0.220. Rater agreement allows about 0.744 within speakers against the released total, which is a guide and not a bound: the released total is a floored median of the five raters rather than

their mean, and the figure is the square root of the within-speaker reliability of that mean times the correlation of the released label with it. A single rater reaches 0.507 against the unrounded median of the other four.

Progress improved both components: HMamba reaches 0.824 pooled and 0.479 within speakers against 0.744 and 0.379 for GOPT, a paired difference of 0.080 and 0.100 with intervals excluding zero, and 0.030 and 0.057 against HierCB with ConPCO. Against that, the noise of the training pipeline is not small: across five seeds of GOPT the pooled total-score correlation spans 0.021, and selecting the epoch on the test set, which the released checkpoint does, buys up to 0.017 on one seed. The seed spread alone exceeds the 0.019 that separates the five best published results, and the selection gain is almost as large, so a difference of that size on this benchmark is not evidence of a better evaluator. The five are MuFFIN, HMamba, HierGAT, HiPPO and 3MH (Yan et al., 2025a; Chao & Chen, 2025; Yan & Chen, 2024; Yan et al., 2025b; Chao et al., 2023), each a mean over five runs, HierGAT's as MuFFIN's comparison table prints it and HiPPO's for its variant trained without the curricular strategy and the CONO regularizer. Table 70 states the checkpoint-selection regime of every row.

The released label, and the word and phone levels. The released utterance label is a median of five experts, floored for the total score, so within one speaker it takes a median of 3 distinct values over twenty utterances. The corpus releases five per-rater scores for each of the four labels, the total among them and unrounded, so every aspect has an unrounded five-rater mean to compare against. We verified the rule on all 5000 utterances: the three aspects are the median of the five scores and the total is that median floored, which the floor changes on 3630 of them. Two things separate the released label from an unrounded five-rater mean, the aggregation rule and, on the total alone, the floor, and the two must be told apart before either is blamed. On the three aspects the released label *is* the unrounded median, exactly, on every one of the 5000 utterances, so there is no rounding step at all and the whole difference is the median against the mean: within speakers the two agree at 0.861 on accuracy, 0.834 on fluency and 0.814 on prosody. On the total both steps act, the floor at 0.920 within speakers against 0.974 pooled and the aggregation at 0.904 against 0.970. Both steps cost within-speaker signal and neither costs much between-speaker signal. The shortcut is therefore *larger* on the cleaner label, on every one of the four: for the total the between-speaker share of the unrounded mean is 0.75 against 0.70 for the released label, with the oracle at 0.850 against 0.816; across the four labels the gap runs from 0.05 to 0.09. Reporting on the released label is the conservative choice.

The corpus also releases word-level and phone-level labels, which is the granularity pronunciation feedback uses, and there the picture changes. Clustered by speaker, word accuracy has a between-cluster share of 0.21 over 15967 words and an oracle of 0.449, word stress 0.12 and 0.331, and phone accuracy 0.20 and 0.439 over 47369 phones, against 70% and 0.816 for the total score at the utterance level. Clustered by utterance instead, the shares rise to 0.35 for word accuracy and 0.45 for stress, while the phone's own identity carries 0.03 over 65 phone types. The speaker shortcut that dominates the utterance-level leaderboard is three to four times weaker at the level at which feedback is given, where the utterance carries the larger share; the oracle does not follow the share there, because a word's twenty-odd utterance neighbours are fewer than a speaker's several hundred words. That level is a three-level design, phones within utterances within speakers, which the two-level framework of Section 3 does not cover.

D Estimators and Their Checks

The corollary in angle form, and what the interval for r_B covers. Writing $\eta_p = \cos\alpha$ and $\eta_y = \cos\beta$ turns Equation 1 into a cosine addition and the threshold of Corollary 1 into $\cos(\alpha+\beta)$, so a reader who knows η_y and r_T but not η_p faces an unrestricted r_W for some admissible evaluator exactly when $|r_T| < \eta_y$, and only for evaluators with $\eta_p \geq \cos(\arccos|r_T| - \beta)$. The inequality is strict: $\eta_p\eta_y - D$ increases in η_p and reaches η_y only at $\eta_p = 1$, where the within-cluster part of p is zero and r_W is undefined, so at $|r_T| = \eta_y$ the interval of every admissible evaluator falls short of $[-1, 1]$ at one end: at $r_T = \eta_y$ its lower end is above -1, at $r_T = \eta_y = 0.5$ equal to $\eta_y(1-\eta_y^2)^{-1/2}\sqrt{(1-\eta_p)/(1+\eta_p)} > 0$, which is 0.132 at $\eta_p = 0.9$ and 0.013 at $\eta_p = 0.999$, and at $r_T = -\eta_y$ it is the upper end that is below 1, by the same amount, while the other end can reach the boundary. The sample r_B is attenuated, a matter of estimand and not of method: it covers the latent between-cluster correlation in 87% of runs at 125 clusters against 92% for the estimand the statistic

does have, the correlation of the cluster means a benchmark observed, which carry within-cluster noise at rate $1/n_c$. A between-cluster correlation is therefore a statement about those observed means and not about the speakers or systems behind them.

Reliability with a fixed rater panel. Let $X \in \mathbb{R}^{N \times K}$ hold the ratings of K raters. The pooled consistency reliability of the rater mean is $\rho_T = (\mathrm{MSR} - \mathrm{MSE})/\mathrm{MSR}$ from the two-way analysis of variance. For the within-cluster reliability, each column of X is centred by its cluster means, which gives $\tilde{X}$ with row mean square $\mathrm{MSR}_W = K \sum_i \tilde{\bar{x}}_{i\cdot}^2/(N - C)$ and residual mean square $\mathrm{MSE}_W = \sum_{ij}(\tilde{x}_{ij} - \tilde{\bar{x}}_{i\cdot})^2/((N - C)(K - 1))$. Under a model with item true scores, rater main effects, cluster by rater interactions, and independent rating errors, $\mathbb{E}[\mathrm{MSR}_W] = K s_W^2 + \sigma_e^2$ and $\mathbb{E}[\mathrm{MSE}_W] = \sigma_e^2$, where s_W^2 is the within-cluster variance of true scores, so $\rho_W = (\mathrm{MSR}_W - \mathrm{MSE}_W)/\mathrm{MSR}_W$.

Which reliability estimator a release licenses. The fixed-panel estimator above needs the column index of a rating table to mean the same rater on every item, which a release that withholds rater identities does not promise. We apply one rule rather than choosing per benchmark. Where real rater identities are released, the two-way model is fitted on them. Where only column positions are released, the fixed-panel estimator is used if and only if the position test of Appendix Y passes, and the one-way estimator otherwise. Applied to the 2 benchmarks that pass and the 5 that fail, the rule selects the fixed-panel estimator for speechocean762 and for the SummEval expert ratings, where position carries a main effect of $F = 439$ on $(2, 3198)$ for coherence and the offset profile reproduces across disjoint halves of the systems, and the one-way estimator for FRANK, RealSumm and HelpSteer2, where it does not. Two negative controls fix the rule's calibration: the SummEval crowd ratings, which the release states come from different workers per item, and HANNA's within-story row index, which counts rows over a varying triple. Both are read by the same code and both fail, the crowd ratings at $F \leq 1.3$. Reporting both estimators is cheap and we do: on SummEval coherence the attainable within-system correlation is 0.840 under the one-way estimator and 0.897 under the fixed-panel one. No claim in the main text turns on which is used, since the SummEval rows of Table 3 are label-only statistics that do not involve a reliability estimate. One caution about naming: the within-cluster term centres each rater inside each cluster, which removes the cluster by rater interaction as well as the rater main effect, so it is not simply a leniency correction. On speechocean762 fluency the gap between the one-way and fixed-panel estimates decomposes into 0.54 from the rater main effect and 0.24 from the interaction, and on clusters of two items the interaction has nearly as many parameters as cells, which is why we do not use the estimator there.

Reliability with varying raters. When item i receives k_i ratings from varying raters, the rating error variance σ_e^2 is the pooled within-item variance, the error variance of the item mean is σ_e^2/k_i, and the error variance of its within-cluster deviation is $\sigma_e^2/k_i\,(1 - 2/n_c) + n_c^{-2} \sum_{j \in c} \sigma_e^2/k_j$. The within-cluster reliability is one minus the summed error variance over the summed squared deviations. Rater leniency is part of the error in this design.

Simulation checks. The released toolkit includes tests on simulated data, with 125 clusters of 20 items unless stated otherwise and 5 to 187 clusters for the identity and reliability checks. The decomposition identity holds to 10^{-10} for balanced and unbalanced clusters. The in-sample share exceeds the analysis-of-variance intraclass correlation by $(1 - \rho)/n$, and that estimator's mean is within 0.01 of ρ at the 125 clusters the test uses. That is a statement about this design and not about unbiasedness: the estimator is a ratio whose numerator and denominator are separately unbiased, and the ratio is biased downward when clusters are few. Over $40{,}000$ draws at $\rho = 0.7$ its mean is 0.662 at nine clusters of twenty items, 0.680 at sixteen clusters of seventeen and 0.698 at 125 clusters of twenty, so the population ceiling $\sqrt{\rho}$ is conservative on the benchmarks here with fewest clusters. The profile-oracle curve matches simulation within 0.012 for $k \in \{1, 3, 10\}$. Cluster-bootstrap 95% intervals lose coverage as clusters become few, and for r_T they are short of nominal even when clusters are many. Over 300 simulated benchmarks per design, drawn with known population values, the interval for r_T covers in 83% of runs at 9 clusters, 90% at 16, 91% at 20 and 92% at 125, and the interval for r_W in 89%, 91%, 91% and 97%. So the rule is not that intervals below twenty clusters are descriptive and the rest exact: every interval on a single benchmark in this paper is read as descriptive, and the claims taken across benchmarks in Section 5 rest on effect sizes with intervals from clusters resampled within each corpus and jointly across a benchmark's configurations (Section 4), whose coverage for the paired utility difference is measured in Appendix AA, and on

the sign over benchmarks, rather than on any single-benchmark interval. Both reliability estimators recover the simulated within-cluster reliability within 0.03.

The reporting protocol. Release the label-only statistics and rater identities: the between-cluster share and intraclass correlation of every label under every cluster variable its item table names, with the population ceiling and the reliability of both components, since the share measures how much is at stake and hiding rater identity charges leniency to error. Report r_T, r_W and r_B for every evaluator with cluster-bootstrap intervals and, wherever the decision is taken inside a cluster, a statistic grouped on that cluster, with the evaluator each names, the decision it is for and the distinct clusters it was read from; where the decision is known and the clusters are many, its own statistic on held-out clusters beside the grouped one (Appendix K). Compare evaluators with a paired cluster bootstrap and report the minimum detectable difference; the interval under-covers at every cluster count simulated (Appendix Z). Beside any evaluator selected from few labelled clusters report the utility of the selected evaluator on clusters the rule did not use, with its interval and its leave-one-out over tasks or configurations, and beside it the regret pooling's pick leaves and the gain the set could detect (Appendix P), since at eight labelled units the sign of a grouped statistic's gain over pooling differs by panel and rests on single tasks (Sections 5 and 6). State how the labels were spent, the clusters, the candidates each labelled cluster holds and the clusters that still ask a question of a within-cluster rule: at eight clusters removing candidates lowers the grouped mean's gain over pooling on the exploratory configurations, and with two candidates a cluster every within-cluster rule selects worse than pooling there (Appendix N). The panels of this paper give no threshold in candidates, no rule for few-valued labels and no statistic that is safe against pooling at every budget: the explanation that few values separate a panel poorly was withdrawn when its predicted sign change failed, and the within-cluster Pearson coefficient, whose best-of-n intervals with eight labelled clusters or half never lie wholly below zero on the panels as released (Appendix Q), has set means that hide configurations more than half a point below pooling (Table 12), loses on a few cells under the pairwise choice and at 32 labelled segments, and picks worse than pooling with eight labelled clusters of two candidates each.

Running the protocol. The label-only statistics, the per-evaluator components and the paired comparisons of the reporting protocol above are one command on a table with one row per item:

```
python -m clustereval items.csv --cluster speaker --label total \
       --pred gopt --pred wavlm
```

prints the number of clusters and their sizes, the between-cluster share, the intraclass correlation, the ceiling with its cluster-bootstrap interval, the leave-one-out oracle, the profile curve, then one line per evaluator with r_T, r_W, r_B and the interval on r_W, and one line per pair of evaluators with the paired difference, its interval, its p-value and the minimum detectable difference. It warns when there are fewer than twenty clusters. Adding `--group doc_id --method spearman` switches to a per-document protocol and reports the same quantities inside the complete block, and `--json` writes them to a file. On the speechocean762 table this reproduces the ceiling 0.827, the oracle 0.816 and the GOPT against WavLM ridge comparison of Appendix C; on SummEval it reproduces the coherence row of Table 3.

Two criteria that cannot choose a grouping. In a simulation with 120 clusters of 12 items and a between-cluster share of 0.6, a family of evaluators with exactly zero within-cluster validity reaches a split-half ranking consistency of 0.91 under the pooled coefficient against 0.01 under the within-cluster one, and separates 84% of evaluator pairs against 8%. Both arms use the same folds, the same half-splits and the same aggregation mask.

A trap in fitting item-blind predictors. Proposition 3's predictor may be read off the labels, but a *fitted* model must not take it as a feature, since $(S_c - y_i)/(n_c - 1)$ determines y_i given the cluster sum; every item-blind predictor we fit therefore uses split-half partitions of each cluster, the profile of one half predicting the labels of the other.

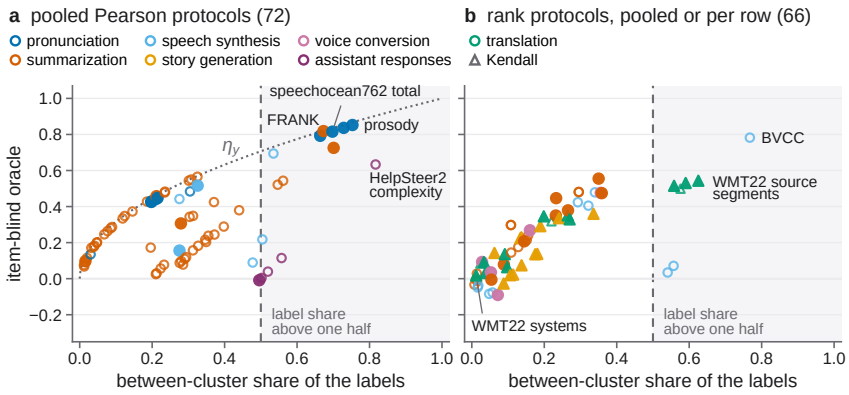


Figure 4: **The census.** Every configuration of the census, at the between-cluster share of its labels and the correlation of the leave-one-out label oracle of Proposition 3 under its own protocol. Filled markers are the rows of Table 3. The left panel holds the configurations whose protocol is Pearson pooled over items, which is where the ceiling η_y of Proposition 1 on a cluster-constant predictor and Corollary 1 apply; the oracle is not cluster-constant and lies below the curve because its within-cluster term is negative. Every per-row configuration of the census is a rank protocol and sits in the right panel with the pooled rank ones, where the label-only maximum is the different quantity of Propositions 4 and 5 and no curve is drawn. The dashed line marks a label share of one half, the corollary's condition for an evaluator whose between-cluster share matches the label's; one with a higher share can be unidentified to the left of it too.

E THE BETWEEN-CLUSTER COMPONENT, MEASURED IN FULL

speechocean762 (Zhang et al., 2021) holds 5000 English utterances read by 250 Mandarin-speaking learners, each test speaker reading 20 sentences, each scored by five experts for accuracy, completeness, fluency and prosody and, in the release, for a total, and published systems report Pearson over the 2500 test utterances. Leaving out completeness, which is at its maximum on 99.5% of utterances, between 66% and 75% of that variation lies between speakers, every aspect's best published result lies inside the interval of its population ceiling, and a regression on WavLM features (Chen et al., 2022) of the speaker's *other* utterances reaches 0.737 with no label at test time, against 0.735 for the same regression on the scored utterance under the same half-partition (0.739 on the whole test set, Table 68); of four published systems we reproduce (Gong et al., 2022; Chen et al., 2024b; Yan et al., 2025c; Chao & Chen, 2025), two lie within 0.007 of that regression pooled while the within-speaker correlations of the five run 0.379 to 0.271. The systems that grade learners in production are commercial services behind an interface that no benchmark scores; we score two, Microsoft Azure AI Speech on all 2500 test utterances and iFlytek on 519, called in September 2026 on the released audio and reference text (Appendix X). From its pooled correlation and the two shares, the between-speaker share of Azure's own outputs, 0.67, and the label share the benchmark publishes, Corollary 1 restricts its within-speaker correlation to $[0.362, 1.000]$, and the labels bear that out: Azure reaches 0.796 pooled on the total, below the leave-one-out oracle at 0.816 and below the best published system, while its within-speaker correlation is 0.501 $[0.401, 0.590]$, above three of the four reproduced systems by paired intervals read at the level such an interval has at 125 clusters (Appendix Z) and +0.022 against the strongest, which is not a separation; iFlytek agrees, 0.658 pooled and 0.410 within. The pooled leaderboard puts a deployed service below a label oracle that reads no audio of the scored utterance, while the component governing the feedback it exists to give puts it among the strongest evaluators here. A third arm changes what an evaluator is *told*: an LLM judge rating 444 SummEval summaries, told the source system is strong or weak, moves the between-system component +0.126 $[+0.022, +0.303]$ more than an identical repeat does and nothing within systems (Appendix X).

Across benchmarks, Table 3 in Appendix F gives ten rows of the label-only statistics and Table 72 all 28, each under the benchmark's own protocol: how much of the published coefficient is reachable without reading the item, a fact about the coefficient and not a verdict on any evaluator, since where the alternatives differ by system, system identity is legitimate information. On 16 of the 22 rows whose published number's item set, aggregation and coefficient were confirmed to be ours, the

Table 3: Label-only statistics, ten rows of the 28 in Table 72, each on the item set and in the coefficient its benchmark's published numbers use. Share is the between-cluster share of the label variance, Bound what no cluster-constant predictor exceeds under that protocol, an exact maximum except where marked §, and Oracle the leave-one-out label oracle, which reads the cluster's other labels and never the item it scores, its bracketed interval the 2.5 and 97.5 percentiles of a cluster bootstrap of 400 resamples (300 where raters are crossed with items). Last column: best published result, then the oracle over it where item set, aggregation and coefficient match (‡ all but a confirmed item set, † another protocol). C is starred below twenty. Appendix F documents every row.

Benchmark	Label	Cluster	Protocol	C	Share	Bound	Oracle [boot. 2.5–97.5%]	Best / ratio
speechocean762	total	speaker	utt. Pearson	125	0.70	0.835	0.816 [.77, .85]	0.830 / 0.98
speechocean762	prosodic	speaker	utt. Pearson	125	0.75	0.868	0.852 [.81, .88]	0.843 / 1.01
BVCC	mos	system	utt. Spearman	187	0.77	0.876§	0.782 [.70, .82]	0.897 / 0.87
SOMOS	mos	system	utt. Spearman	201	0.34	0.577§	0.479 [.39, .52]	0.681‡ / (0.70)
VCC2018	mos	system	utt. Spearman	26	0.36	0.564§	0.476 [.25, .58]	0.718 / 0.66
SummEval	coherence	system	doc. Spearman	16*	0.35	0.567	0.555 [.52, .59]	0.652 / 0.85
RealSumm	LitePyramid	system	doc. Spearman	24	0.11	0.309	0.298 [.23, .36]	0.53‡ / (0.56)
FRANK	factuality	system	pooled Pearson†	9*	0.67	0.820	0.819 [.63, .88]	0.47†
WMT22 en-de	mqm	system	seg. Kendall	13*	0.01	0.105§	0.018 [.01, .03]	0.369 / 0.05
WMT22 en-de	mqm	segment	seg. Kendall	1315	0.59	0.709§	0.531 [.51, .55]	0.369 / 1.44

oracle reaches at least two thirds of the published result, from 5% on the WMT22 en-de system row, a negative control, to 144% on its source-segment row. A census puts those rows in proportion (Figure 4): crossing every numeric human label of every benchmark we normalise with every cluster variable its item table names gives 201 configurations over the 16 benchmarks and parts of Table 63, 12 benchmarks by the count of the text, 138 of them estimable; the median configuration is not in danger, at a share of 0.21, and 21 put more than half the label variation between clusters. The shortcut belongs to a cluster variable and not to a benchmark: on WMT22 MQM an oracle that knows only the source segment reaches 0.515 to 0.543 on the three pairs, above every submitted metric, while one constant within systems reaches 0.018, a negative control.

F How Each Row of Table 3 Is Built

The curve of Proposition 2 predicts the measured split-half oracle within 0.019 over the 15 configurations whose intraclass correlation is not negligible, which are the ones on which it is not degenerate. Every row uses one configuration of its benchmark, listed in Table 4: the item set the published numbers are computed on, the coefficient those numbers use, and the grouping they use. Where the benchmark's papers report per-label results, the comparison value is the strongest published value of that label, so the column is a maximum over papers and no single system attains all of it. Every value was read from the primary source rather than from a table that transcribes it, and a ratio is printed only where the item set, the aggregation and the coefficient were confirmed to match ours. Two SummEval rows needed that check. HD-Eval (Liu et al., 2024) prints 0.657 on coherence and 0.599 on relevance, higher than the values used here, and Hu et al. (2024) transcribes its column into a summary-level table; but HD-Eval holds out half of each dataset, fits a regressor on the other half and reports one pooled coefficient over the test half. Reproducing its own expert-against-expert table, three expert pairs by four aspects by two coefficients, settles the aggregation: over 200 random halves, pooling reproduces 23 of the 24 printed values to within two standard deviations and the last at 2.1, while per-document averaging misses 15 of them by more than two and up to 6.4 (`scripts/text/hdeval_table9.py`), so it is on a different protocol. The same-protocol maxima are Cho et al. (2026) for coherence and Kim et al. (2024) for relevance, both of which report summary-level correlations, which the protocol they follow computes per document, and both leave our ratio above two thirds. Three further rows need their restriction stated. The RealSumm value is the strongest metric that reads no human content units at inference, quoted to the two decimals its source prints (Zhang & Bansal, 2021, Tables 1 and 4); the semi-automatic Lite2 Pyramid of its Table 1 reaches 0.60 by reading the benchmark's own units for the summary it scores, which is a different information condition, and the fully automatic value is itself a five-fold cross-validated average with the metric fitted in domain, against 0.47 under a split by systems and 0.50 for the strongest metric that is not fitted at all. Its table uses the benchmark's 25 columns where our row uses 24, which is why that row is marked as matching on aggregation and coefficient but not on the item

set. The release is the reason: its `abs.pkl` entry for BART is a byte copy of the `ext.pkl` one, so the abstractive BART system carries no official LitePyramid label and the published 25-column table holds the extractive BART column twice. Rebuilding that table by duplicating the column moves the oracle from 0.298 to 0.304, a ratio of 0.57 against 0.56, so the caveat acts in the direction that helps us and neither figure is near two thirds. On the mixed set no released metric score reaches 0.50. The VCC2018 value is the 0.718 of Tseng et al. (2021), on the released 13,580 by 3000 by 4000 split of the 20,580-utterance set that the row uses: the paper gives the sizes, and its official recipe ships this list; Wang et al. (2023c) report 0.721 on a split of the same sizes whose list they do not name, and a higher figure on a differently drawn split of a larger corpus is not comparable either. The SOMOS value is marked like the RealSumm one for a reason of that kind: Vioni et al. (2023) give 3000 as the size of both their validation and test sets and do not say which one their Table 3 uses. Two rows carry a dagger and two a double dagger. A cluster that holds a single item has no cluster mean that is not its own label, so the oracle, whose predictor may not read the item it scores, falls back to the global leave-one-out mean there; the Bound, which is an in-sample maximum and may read it, does not. This affects only BVCC, where 45 of its 187 systems contribute one test utterance each, and treating the two alike would understate that row's bound by 0.0308.

FRANK is the one benchmark whose own protocol already removes the shortcut. Its evaluation script reports a partial Pearson correlation with system identity partialled out (Pagnoni et al., 2021), and a system-constant predictor has zero residual in that regression, so its partial correlation is a ratio with a zero numerator and a zero denominator and is undefined, whatever the labels are; implementations that return zero there do so by convention. The row therefore reports the plain pooled Pearson correlation, where the oracle reaches 0.819, and prints no ratio against the partialled value 0.47. That value is the partial Pearson correlation of FactKB (Feng et al., 2023), not the partial Spearman correlation of 0.52 in the neighbouring column of the same table; FactKB writes ρ for Pearson and r for Spearman. Its nine systems are nested in two source corpora, and the census separates the two: the between-cluster share of factuality is 0.67 by system over both corpora but 0.59 by corpus alone, 0.31 by system inside CNN/DailyMail and 0.02 inside XSum. The benchmark's own protocol partials out system and dataset together for exactly this reason.

HANNA reports story-level Kendall correlations for each of its six criteria and for their average (Chhun et al., 2022); the 0.27 for chrF in its main figure is the average, and its Table 8 gives the five strongest metrics of each criterion under each of three coefficients. The strongest Kendall value of each criterion is the comparison value of its HANNA row: recomputed under the same per-prompt protocol, with the human stories excluded and the absolute value taken as HANNA takes it, the released metric scores reproduce all six to four decimals, ROUGE-WE-3 recall on coherence at 0.2529 against the printed 25.29% and chrF at 0.2461 against 24.61%, so the item set, the aggregation and the coefficient are confirmed. The strongest of the 92 evaluators in the benchmark's release, the OrcaPlatypus column of the rating file released with Chhun et al. (2024), a rater that paper does not describe, reaches 0.301 on coherence, above the oracle and above every story-level Kendall value HANNA prints for coherence. The coherence row also has the weakest support of any in the table: its within-system reliability estimate for coherence is not positive under the one-way estimator, at -1.08. That is an artefact of the estimator and not a property of the labels: HANNA releases real annotator identities, and a two-way model fitted on them gives 0.32, an attainable within-system correlation of 0.57 rather than zero.

TTSDS2 reports system-level Spearman correlations computed separately for each of twelve domain and rating-type cells (Minixhofer et al., 2026) and publishes no utterance-level result, so the row's utterance-level Pearson correlation is not a protocol this benchmark uses and has no comparison value.

The WMT22 source-segment rows use the same published comparison value as the system rows, because both are the flattened segment-level protocol of the shared task over the same items. The oracle in those rows is not a metric: it predicts a translation from the human scores of the other systems' translations of the same source segment. Human reference translations and the systems the shared task marks as outliers are excluded, which is the configuration that reproduces the published Kendall correlations of Table 11 of Freitag et al. (2022) to three decimals.

Table 4: Configuration of every row. Item set is the set of items the row is computed on, chosen to match the set the published numbers use. Published result names the value in the last two columns of Tables 3 and 72 and its source.

Benchmark	Label	Item set	Published result
speechocean762	accuracy	official test split, 2500 utterances	utterance Pearson, HMamba (Chao & Chen, 2025)
speechocean762	fluency	official test split, 2500 utterances	utterance Pearson, HMamba (Chao & Chen, 2025)
speechocean762	prosodic	official test split, 2500 utterances	utterance Pearson, HMamba (Chao & Chen, 2025)
speechocean762	total	official test split, 2500 utterances	utterance Pearson, MuFFIN (Yan et al., 2025a)
BVCC	mos	official test split, 1066 utterances	utterance Spearman, UTMOS (Saeki et al., 2022), team T17 in Table 5 of Huang et al. (2022a)
SOMOS	mos	official test split, 3000 utterances	utterance Spearman, Vioni et al. (2023) Table 3, a split of the same size whose list the source does not name‡
VCC2018	mos	LDNet test split, 4000 utterances	utterance Spearman, Tseng et al. (2021) Table 2(a)
TTSDS2	mos	all 466 rated utterances, four domains pooled	no published result under this protocol
SummEval	coherence	expert set, 100 documents by 16 systems	per-document Spearman, BinEval (Cho et al., 2026)
SummEval	consistency	expert set, 100 documents by 16 systems	per-document Spearman, BinEval (Cho et al., 2026)
SummEval	fluency	expert set, 100 documents by 16 systems	per-document Spearman, Themis (Hu et al., 2024)
SummEval	relevance	expert set, 100 documents by 16 systems	per-document Spearman, DEBATE (Kim et al., 2024)
RealSumm	litepyramid recall	100 documents by 24 systems, the one system with no LitePyramid label dropped	per-document Spearman, Lite3 Pyramid (Zhang & Bansal, 2021), strongest metric using no human content units, same aggregation and coefficient on a different item set‡
HANNA	coherence	96 prompts by 10 systems, human stories excluded	per-prompt Kendall, ROUGE-WE-3 recall, Table 8 of Chhun et al. (2022)
HANNA	relevance	96 prompts by 10 systems, human stories excluded	per-prompt Kendall, SUPERT-SS, Table 8 of Chhun et al. (2022)
HANNA	empathy	96 prompts by 10 systems, human stories excluded	per-prompt Kendall, ROUGE-WE-3 recall, Table 8 of Chhun et al. (2022)
HANNA	surprise	96 prompts by 10 systems, human stories excluded	per-prompt Kendall, chrF, Table 8 of Chhun et al. (2022)
HANNA	engagement	96 prompts by 10 systems, human stories excluded	per-prompt Kendall, chrF, Table 8 of Chhun et al. (2022)
HANNA	complexity	96 prompts by 10 systems, human stories excluded	per-prompt Kendall, chrF, Table 8 of Chhun et al. (2022)
FRANK	factuality	released annotations, 2246 summaries by 9 systems over 499 articles, both source datasets	partial Pearson on the 1575-summary test split, FactKB (Feng et al., 2023) Table 3, a different protocol, so no ratio†
HelpSteer2	helpfulness	validation split, 1038 responses	no published result under this protocol
TL;DR axis	overall	Reddit TL;DR axis ratings, 7682 summaries by 13 policies	no published result under this protocol
WMT22 en-de	mqm	MQM segments, 1315 segments by 13 systems, human references and outlier systems excluded	segment Kendall, UniTE, Table 11 of Freitag et al. (2022)
WMT22 en-de	mqm, source clusters	MQM segments, 1315 segments by 13 systems, human references and outlier systems excluded	segment Kendall, UniTE, Table 11 of Freitag et al. (2022)
WMT22 zh-en	mqm	MQM segments, 1875 segments by 14 systems, human references and outlier systems excluded	segment Kendall, COMET-22, Table 11 of Freitag et al. (2022)
WMT22 zh-en	mqm, source clusters	MQM segments, 1875 segments by 14 systems, human references and outlier systems excluded	segment Kendall, COMET-22, Table 11 of Freitag et al. (2022)
WMT22 en-ru	mqm	MQM segments, 1315 segments by 15 systems, human references and outlier systems excluded	segment Kendall, MetricX, Table 11 of Freitag et al. (2022)
WMT22 en-ru	mqm, source clusters	MQM segments, 1315 segments by 15 systems, human references and outlier systems excluded	segment Kendall, MetricX, Table 11 of Freitag et al. (2022)

G What Each Benchmark Headlines

The selection experiments of Section 5 and Appendix H score each benchmark's own statistic beside the pooled and within-cluster coefficients, so that statistic has to be identified from the primary source rather than assumed. The result is the census below, which is also the answer to what this paper asks anyone to change. Three of the six publish a statistic computed inside the source document, prompt or segment and three do not; FRANK is the one whose published statistic groups on the other factor, the system, and which therefore already reports a within-cluster coefficient for a decision taken inside a system and not for one taken inside a document.

Machine translation. The WMT22 metrics task reports segment-level Kendall's τ_b under what it calls *none* averaging, flattening the system by segment matrix into one vector before correlating (Freitag et al., 2022); it computes the by-segment and by-system averages as well, but publishes no table of them, using them only inside a weighted rank over 201 tasks. WMT23 ranked metrics by a weighted average of ten tasks: system-level pairwise accuracy and, per language pair, system-level Pearson, segment-level Pearson on the flattened system-by-segment matrix, and the pairwise ranking accuracy with tie calibration, the acc^*_{eq} of Deutsch et al. (2023), computed within each source segment and averaged over segments (Freitag et al., 2023), so that half its segment-level weight was pooled Pearson and the benchmark's own statistic scored here is its grouped half; WMT24 kept acc^*_{eq} as the only segment-level statistic beside a system-level soft pairwise accuracy (Thompson et al., 2024), dropped Pearson entirely, and gave as the reason that metrics are used "to rank a set of translations for the same source segment" in minimum-Bayes-risk decoding and quality-estimation reranking (Freitag et al., 2024). WMT25 keeps both unchanged (Lavie et al., 2025), so what is scored here as this benchmark's own is its segment-level protocol, the one whose decision is taken inside a source segment, and not the task's only headline. A per-segment Kendall average was not WMT's headline statistic from 2022 to 2025, entering only the 2022 weighted rank as some of its 201 tasks; it is scored here as a control rather than as a protocol, on the same grouping as acc^*_{eq} and without the tie calibration, because that statistic changes the grouping and the coefficient at once and the two effects are otherwise not separable. Translation is therefore the one field here that has already moved its headline statistic inside the source segment. The statistic scored as this benchmark's own is accordingly the current one and not the 2022 rule, which the task has replaced; the 2022 rule is scored beside it, and the two rules give the WMT row of Table 65 the same own column, 25.8%; both are printed in the release. The comparison belongs on the cells whose decision is taken inside a source segment, which is the grouping all three rules use. Our sweep maximises one threshold over the pooled pairs where WMT averages a per-segment accuracy under one threshold; the two agree because every system translates every segment here, so the segments hold equally many pairs, and the code refuses to run where they do not. On those 6 cells the tie-calibrated accuracy errs on 25.8% of evaluator pairs, against 21.0% for the Kendall control on the same grouping, 25.8% for the flattened 2022 rule, 28.0% for the pooled Pearson coefficient, 20.9% for the within-cluster one and 14.3% for the grouped mean, with 11.2% for running the decision itself. Two readings follow, and the first has to be scoped. Tie calibration solves the problem it was introduced for, since MQM scores tie heavily inside a segment, no Kendall variant credits a correctly predicted tie, and regression metrics almost never predict one (Deutsch et al., 2023). As a rule for *choosing between metrics* it sits behind a plain Kendall mean on the same grouping, and well behind the grouped Spearman mean, so on these cells and *against a decision that counts only pairs the humans separate* what pays is the grouping and not the tie calibration. Against a decision that pays for recognising a tie the order reverses and the same statistic is the best rule of the six on every benchmark here, which Appendix I measures; a comparison of meta-evaluation statistics is a comparison relative to a utility, and this paper's headline utility does not reward identifying an equivalence. And grouping on the decision's cluster is necessary and not sufficient: on the 6 system-clustered cells, where the segment grouping is not the decision's, the same two rules run 37.9% and 29.7% against 24.9% for a mean grouped on the system. Twelve cells of one benchmark family is descriptive and is reported as such. Our implementation sweeps the tie threshold exactly rather than over a grid and agrees with a direct enumeration to 0e+00 on 40 random instances (`scripts/decision/selection_rules.py`).

Summarization and story generation. FRANK partials a categorical variable encoding the system and the dataset out of both the metric and the human score and correlates the residuals, pooled over all summaries (Pagnoni et al., 2021); partialling out system identity is centring within systems, so on a system grouping that statistic is the within-cluster correlation in residual form, which the experiment confirms by reproducing it to machine precision. RealSumm reports a system-level and a per-document correlation side by side and finds that the two can disagree (Bhandari et al., 2020); on its system-clustered configuration the per-document rule orders evaluators for a within-system decision at a rank correlation of -0.05 across the sixteen metrics, which is why Table 65 has it erring above chance there. HANNA's *story-level* correlation is computed across the ten system outputs of one prompt and averaged over the 96 prompts (Chhun et al., 2022), which is a per-prompt grouping despite the name; the same authors' later work drops it for a pooled coefficient on the ground that stories from one prompt, unlike translations of one sentence, need not be similar (Chhun et al., 2024). SummEval is the exception in its own field: its paper correlates metrics with the expert ratings only at the system level, by Kendall's τ, and reports no item-level metric correlation (Fabbri et al., 2021),

Table 5: **The utility of the selected evaluator on clusters no rule used.** Every rule is computed on half the clusters of a configuration and selects one evaluator, scored on the other half (Section 4). Cfg: configurations; Ev.: evaluators in the panel, 17 or 18 on WMT22 by language pair. Columns r_T to own: the payoff of the selected evaluator there, for the pooled and within-cluster coefficients, the grouped mean, the payoff's decision statistic computed on the selection half (dec: each evaluator's best-of-n recovery, or its pairwise accuracy, on the labelled clusters) and the benchmark's own statistic; best: the best evaluator on the evaluation half. grp$-r_T$ and grp$-$own: the utility gain of the grouped mean's pick over pooling's and the own statistic's, paired draw by draw, the estimate ± 1.96 bootstrap standard deviations, descriptive and not tests; the coverage of grp$-r_T$ is measured in Appendix AA and that of grp$-$own is not. Err: the share of evaluator pairs on which the rule selects the one that is worse on the evaluation half. Every configuration: Table 64.

			Payoff of the pick ↑						Utility gain, points		Err, % ↓	
Benchmark	Cfg	Ev.	r_T	r_W	grp	dec	own	best	grp$-r_T$	grp$-$own	r_T	grp
best of n, recovery %												
FRANK	1	11	43.4	49.8	49.8	48.7	47.6	49.9	+6.4 [−4.1, +17.0]	+2.2 [−4.5, +8.8]	46.8	35.6
HANNA	6	19	28.9	31.7	32.5	32.3	32.1	40.1	+3.7 [−0.9, +8.2]	+0.4 [−0.5, +1.3]	18.8	16.7
RealSumm	1	16	38.3	40.0	40.7	33.7	40.7	44.5	+2.4 [−4.0, +8.7]	+0.0 [+0.0, +0.0]	44.1	28.0
SummEval	4	16	74.2	73.9	72.0	74.1	72.0	78.1	−2.2 [−5.0, +0.6]	+0.0 [−0.0, +0.0]	21.0	22.7
WMT22 MQM	3	17–18	46.1	46.7	46.5	44.9	46.8	48.9	+0.4 [−1.6, +2.4]	−0.3 [−1.9, +1.3]	35.1	19.3
speechocean762	4	9	49.0	48.4	50.9	45.5	49.0	54.3	+1.9 [−1.9, +5.7]	+1.9 [−1.9, +5.7]	31.6	27.9
All primary	19	–	46.6	47.9	48.3	46.8	47.7	53.2	+1.6 [−0.2, +3.5]	+0.6 [−0.3, +1.6]	27.4	22.3
All but WMT	16	–	46.7	48.1	48.6	47.2	47.8	54.0	+1.8 [−0.3, +4.0]	+0.8 [−0.3, +1.8]	25.9	22.9
pairwise choice, accuracy %												
FRANK	1	11	73.6	76.8	76.8	76.8	75.7	76.8	+3.2 [−1.2, +7.6]	+1.1 [−2.0, +4.2]	29.5	15.9
HANNA	6	19	63.5	64.9	64.8	64.9	64.7	66.0	+1.3 [+0.0, +2.5]	+0.1 [−0.2, +0.5]	16.7	10.7
RealSumm	1	16	69.7	70.7	70.5	70.5	70.5	71.1	+0.8 [−1.4, +3.0]	+0.0 [+0.0, +0.0]	38.4	13.9
SummEval	4	16	76.6	76.6	77.4	77.4	77.4	77.9	+0.8 [−0.1, +1.7]	+0.0 [−0.0, +0.0]	15.0	12.8
WMT22 MQM	3	17–18	68.6	68.7	68.6	68.6	68.6	68.9	+0.0 [−0.3, +0.4]	+0.0 [−0.3, +0.3]	20.9	9.0
speechocean762	4	9	71.4	70.8	72.1	72.1	71.4	72.3	+0.7 [+0.0, +1.4]	+0.7 [+0.0, +1.4]	21.1	13.2
All primary	19	–	69.6	70.2	70.5	70.6	70.3	71.1	+0.9 [+0.4, +1.4]	+0.3 [−0.0, +0.5]	19.8	11.8
All but WMT	16	–	69.8	70.4	70.9	70.9	70.6	71.5	+1.1 [+0.5, +1.7]	+0.3 [−0.0, +0.6]	19.5	12.4

and the per-document Spearman that its downstream literature reports, and that the comparison values in Table 3 use, is a convention of that literature rather than of the benchmark. We use the downstream convention, since that is what the numbers we compare against are computed under.

Speech. speechocean762's own paper reports phone-level correlation only (Zhang et al., 2021), and its leaderboard reports a pooled utterance-level Pearson (Gong et al., 2022; Chao & Chen, 2025; Yan et al., 2025a). The VoiceMOS challenges and SOMOS report a pooled utterance-level correlation and a correlation over per-system means, with system-level Spearman as the leaderboard's primary metric (Huang et al., 2022a; Maniati et al., 2022). No within-speaker, within-system or within-listener correlation appears in any of them, although listener identity is heavily modelled (Leng et al., 2021; Huang et al., 2022b). The test split of speechocean762 is speaker-disjoint and holds twenty utterances for each of its 125 test speakers, so the within-speaker coefficient has been computable since 2021, and we found no paper that computes it.

H Which Configurations Enter the Decision Experiment

Three nested counts run through the paper and the step between them is a different rule each time, so all three are stated here with the rule that produces them. The census of Table 63 crosses every numeric human label of every normalised benchmark with every column of its item table that names a unit the benchmark samples, which gives 201 configurations, of which 138 have at least five clusters and a typical cluster of more than one item, the two rules of Table 63. The cross-benchmark table takes one configuration per benchmark and label on the item set that benchmark's published numbers use, which gives 28 rows; the step from 201 to 28 is not a filter on results but a choice of the published protocol, and Appendix F names the protocol for each. The decision experiment needs evaluator scores as well as labels, so its criterion is: the release must carry evaluator scores, at least four of them must score every item of the configuration, and the label must vary. Applied to the 19 labels of the benchmarks with a crossed item table, it admits 15 and excludes 4, all of them labels of the TL;DR axis ratings (Stiennon et al., 2020), which release no evaluator scores at all; `scripts/decision/selection_task.py` records the ledger, one line per label with the

Algorithm 1 The held-out estimate of what a rule's pick is worth, for one configuration, one budget and one payoff (`scripts/decision/targets.py`; the selection error by pair as defined below).

Input: clusters $\mathcal{C}$ with labels y; the scores $p^{(1)}, \dots, p^{(K)}$ of K evaluators; two rules A and B (pooled r_T and the grouped mean in the text); a budget b; a payoff U, best of n or the pairwise choice; S splits

for $s = 1, \dots, S$ **do**
  split $\mathcal{C}$ in half at random, whole documents where clusters nest: selection half H, evaluation half M
  draw b distinct clusters $L \subseteq H$, all of H at $b = C/2$, and one random priority π over the evaluators
  for each rule $R \in \{A, B\}$ **do**
    compute R's statistic $a_R^{(k)}$ of every evaluator k on the clusters of L alone
    $\hat{e}_R \leftarrow$ the evaluator with the largest defined $a_R^{(k)}$, ties broken by π ▷ the pick
    $u_R \leftarrow U_M(\hat{e}_R)$, a ratio of sums over the clusters of M ▷ its utility
    $\varepsilon_R \leftarrow$ the share of evaluator pairs decisive on M that a_R orders against U_M, a tie $\frac{1}{2}$ ▷ its error
  end for
  $\Delta_s \leftarrow u_B - u_A, \quad \Omega_s \leftarrow \varepsilon_A - \varepsilon_B$ ▷ paired on the same M
end for

Output: the utility gain $\bar{\Delta} = \frac{1}{S}\sum_s \Delta_s$ and the ordering advantage $\bar{\Omega} = \frac{1}{S}\sum_s \Omega_s$, in points; each interval ± 1.96 standard deviations over 200 corpus-bootstrap replicates of 300 splits, clusters resampled jointly across a benchmark's configurations. On PPE, M is the fixed measurement set of 192 prompts a task, L is drawn from the pool, and $S = 500$.

reason. Each admitted label gives two configurations, clustered by the input and by the system, and each configuration two cells, one per decision score, which is 34 configurations and 68 cells. Every one of them enters every figure quoted; none is selected out. speechocean762 is not crossed and releases no evaluator scores, so its panel, one we assembled and not a sample of deployed systems, is the nine evaluators we reproduced or trained, four published systems, our WavLM regression in two forms, a text-only and a speaker-embedding control, and the cluster-mean-conditioned regression of Appendix AE, the speaker-embedding control built to vary little within a speaker, the kind of evaluator on which pooled and grouped statistics disagree; every one of them scores every utterance on all four released aspect labels, so under the same rule all four enter, accuracy, fluency, prosody and the total, and the fifth label, completeness, is excluded because it equals its maximum on 99.5% of utterances (Appendix Y). Applying the rule rather than taking the total alone adds three configurations and six primary cells, and Table 65 shows each. The two kinds of panel are different objects: on the five text benchmarks the evaluators are the released ones, fixed by the benchmark, and on speech they are a panel we built, which is stated wherever the speech cells are quoted. Held out in this experiment means that the meta-selection is held out, not that every evaluator was blind to the test set when it was trained or selected: Table 70 marks the released GOPT checkpoint as selected on the test set by its authors. `scripts/decision/gopt_sensitivity.py` reruns the 4 speech configurations without it and without both released checkpoints, on the same splits: the pooled coefficient errs on 26.6%, 24.9% and 24.7% and the grouped mean on 20.5%, 18.0% and 16.7% with the full panel, without GOPT and without both, so the speech result does not rest on that checkpoint. Evaluator counts differ between tables by two rules; the second set is not a subset of the first. FRANK releases 15 metric score files. The ranking of Appendix AB takes every file that scores all 2246 labelled summaries, which excludes the entailment metric at 2163 and the question-answering metric at 2242 and keeps 13, among them the article-referenced precision and recall of BERTScore. The decision experiment starts from the 13 metrics listed for the benchmark in the released configuration, which follows the metrics its paper reports and so holds those two metrics and not the article-referenced precision and recall, and then applies its coverage rule, which removes the same two and leaves 11; Table 65 shows that number. Neither set restricts the item set, which is the same 2246 summaries in both; `results/selection_task.json` carries the per-file ledger (`_coverage`) with the count each file scores and the reason it enters or does not.

The decisions, the regret and the statistics, in one place. Algorithm 1 sets out the steps. For a configuration with clusters c of n_c items, labels y and an evaluator's scores p, write $\mathcal{P}_c = \{(i,j) : i < j \in c,\ |y_i - y_j| > 0\}$ for the candidate pairs of c whose labels differ. The pairwise choice is $\sum_c \sum_{(i,j)\in\mathcal{P}_c} s_{ij} \big/ \sum_c |\mathcal{P}_c|$ with $s_{ij} = 1$ if $\operatorname{sgn}(p_i - p_j) = \operatorname{sgn}(y_i - y_j)$, $\frac{1}{2}$ if $p_i = p_j$ and 0 otherwise, so clusters are weighted by their number of decisive pairs. Best of n is $\sum_c (\bar{y}_c^{\,\mathrm{top}(p)} -$

$\bar{y}_c)/\sum_c(\max_{i\in c} y_i - \bar{y}_c)$ over the clusters whose maximum exceeds their mean, where $\bar{y}_c^{\,\mathrm{top}(p)}$ is the mean label of the items p ties at the top of c, so clusters are weighted by the gain available in them. Best of 3 is the same over every triple of a cluster. The three-way decision counts every pair of c, tied labels included, and pays one for calling better, worse or equal as the labels do, the evaluator calling equal when $|p_i - p_j|$ is below its own threshold, which is chosen on the selection half to maximise the same score and then held fixed; acc^*_{eq} on the selection half is that maximum. A rule's selection error over a split is the share of evaluator pairs whose measurement-half decision scores differ by at least 10^{-9} on which the rule selects the one with the lower score, an exact tie in the rule counting one half; the rates are averaged over splits. The held-out script drops a pair on which a rule's statistic is undefined, which at half the clusters is at most 1.14% of pairs for any rule, while the budget scripts, which reach budgets of two clusters, score such a pair as a coin, one half, and the two conventions agree at half the clusters to a tenth of a point on the primary means (Appendix J). Top-1 regret is $d_{\max} - d_{\text{named}}$ on the evaluation half, and its share is that divided by $d_{\max} - \bar{d}$, the span between the best evaluator and the average one. The pooled and within-cluster coefficients are Pearson on all items of the half and on their within-cluster parts; the grouped mean is the equal-weight mean over the half's clusters with a varying label of the per-cluster Spearman correlation, zero where the evaluator is constant; the own statistics are in Appendix G.

Degenerate clusters, ties and the denominator of a grouped mean. A grouped coefficient has to say what it does with a cluster on which it cannot be computed, and the answer decides which clusters each evaluator is scored on. A cluster whose label is constant asks no question and is dropped for every evaluator alike. A cluster whose label varies while an evaluator gives every item of it the same score does ask one, and the evaluator has declined to answer. Dropping it, which is what the implementations in common use do, as Deutsch et al. (2023) note, makes the denominator a property of the evaluator: on 12 of the 34 configurations two released evaluators would then be averaged over different sets of clusters, and the evaluator that sits out the most sits out 41% of the clusters that ask. An evaluator can then raise its score by refusing to order the clusters it would get wrong, which is the NaN problem Deutsch et al. (2023) identify for grouped Kendall's τ and note for Spearman and Pearson as well, each being undefined on a constant vector. We score such a cluster at zero and keep it, so every evaluator is averaged over the same clusters. The choice is not neutral and its size is measured rather than asserted: scored the other way the grouped mean errs on 17.8% of primary pairs against 17.1%, so the convention in common use was making the grouped mean look +0.7 points worse rather than flattering it. Two scripts compute the held-out experiment, one carrying the bootstrap machinery and one the six rules; they draw the same splits from the same seed per configuration and the shared columns agree to 2e-15 on all 68 cells, which the second script checks and records. The remaining conventions are stated once here: an exact tie between two evaluators' selection coefficients is half an error; a pair whose held-out decision scores agree to 10^{-9} is not decisive and is dropped for every rule; a tie in the evaluator's own scores inside a cluster is half a concordant pair in the pairwise score, and an evaluator tying several items at the top of a cluster is credited in best-of-n with their mean label, which is what a uniform tie-break realises in expectation. Taking the first of the tied items instead would credit the evaluator with the order the release file happens to use.

I What the Comparison Depends On

Three things decide the answer of the held-out experiment of Appendix H and are not forced by the data: the criterion a benchmark would use if it never ran a held-out experiment, the utility the decision is scored under, and how far the candidates differ by system. Each is measured on the same configurations and the same 500 splits, so nothing below is a second sample of anything above. These experiments predate the utility estimand of Section 4: they score a rule by its selection error over evaluator pairs on the evaluation half, which they call its cost, the cheaper rule being the one with the smaller error, and Section 7 shows that an advantage in this error, measured on half a benchmark as it is here, says little about the utility of the evaluator a rule selects, so nothing below is evidence about that utility.

Reproducibility and separation against cost. Gao et al. (2025) recommend a grouping by two properties of the statistic: how reproducible the ranking of evaluators is, and how many pairs of evaluators the statistic separates. Both are computable without any decision, and

`scripts/decision/criteria_vs_cost.py` computes them here. Reproducibility is the Spearman correlation between a rule's ranking of the evaluators on half A and its ranking on half B, averaged over splits. Separation is the share of evaluator pairs whose mean gap over splits exceeds 1.96 times the standard deviation of that gap over splits, which is one yardstick applied to every rule because every rule is read off the same halves; a pair with no spread at all counts as separated when its gap is not zero. On the 38 primary cells the pooled coefficient reaches 0.85 and 70%, the within-cluster one 0.88 and 73%, the grouped mean 0.88 and 72%, and each benchmark's own statistic 0.88 and 71%, against selection errors of 23.6%, 19.7%, 17.1% and 20.6%. Cell by cell, pooling is the more reproducible of pooling and the grouped mean on 22 of 38, separates more pairs on 16, and is the cheaper rule on 6; by benchmark it leads on reproducibility on 3 of 6, on separation on 3, and on cost on 0. The rule that leads on a criterion is the cheapest rule on 21 of 38 cells for reproducibility and 16 for separation, against the 10 that naming one of the four rules at random would give. FRANK and WMT are the clean cases and RealSumm the counterexample, where pooling is the least reproducible rule at 0.27 against 0.86 and also the most expensive. Those two are proxies; the criteria as Gao et al. (2025) define them are computed by `scripts/decision/criteria_gao.py` on the same splits. Their ranking consistency splits the inputs in half at random, ranks the evaluators by the measure on each half and takes Kendall's τ between the two rankings, averaged over repeats; the inputs of a primary configuration are its clusters, so the halves are the held-out experiment's own 500 halves and the only change from the proxy is the coefficient. Their discriminative power is the mean over evaluator pairs of the two-sided p-value of the Perm-Both test of Deutsch et al. (2021), which standardises the two evaluators' scores and swaps them item by item with probability one half, at 1000 permutations, computed on the whole configuration as they compute it, smaller meaning more pairs told apart. Four measures are scored. Three are among the twelve that Gao et al. (2025) compare, global Pearson, global Spearman and the input-level Spearman mean, which are the pooled coefficient, its rank form and the grouped mean; the fourth, the within-cluster Pearson coefficient r_W, is not. On the 19 primary configurations the pooled coefficient reaches a consistency of 0.74 and a discriminative power of 0.089 against 0.76 and 0.128 for the grouped mean, so their ordering of global over input-level grouping does not reproduce on consistency and reproduces on discrimination; pooling is the more consistent of the two on 10 configurations and the more discriminating on 15, and the cheaper on 2, by benchmark 3, 5 and 0 of 6. The measure that leads on consistency is the cheapest of the four on 10 configurations and the one that leads on discrimination on 6, so under their own definitions consistency does not separate pooling from the grouped mean and discriminative power favours pooling where pooling is the dearer rule. Appendix D shows the mechanism in simulation, where both criteria rise with the between-cluster share of the labels for an evaluator with no within-cluster validity at all; what this adds is that on the benchmarks in use neither criterion reads the cost.

A utility that pays for recognising a tie. The pairwise choice of Section 4 counts only pairs whose human labels differ, which is what choosing among candidates asks and not what a metric is asked for when an equivalence is itself the judgement. Deutsch et al. (2023) introduce the tie-calibrated accuracy acc^*_{eq} for exactly that, and 33% of the within-cluster pairs of the primary configurations here are tied in the human label, from 11% to 67% by cell. `scripts/decision/tie_aware.py` scores a three-way decision: over every within-cluster pair, ties included, the evaluator must say better, worse or equal, at a tie threshold of its own that is chosen on the selection half, where choosing it is exactly the maximisation acc^*_{eq} performs, and then held fixed and scored on the evaluation half. Two consequences are structural rather than empirical. Under this utility acc^*_{eq} computed on the selection half is not one validation statistic among several: it is the decision itself measured there, the reference rule of Table 65. And an evaluator can score well by calling almost everything a tie wherever ties are common, which is why the tie rate is reported beside every cell. Over the 19 primary configurations the rules run acc^*_{eq} 11.3%, the within-cluster correlation 18.4%, each benchmark's own statistic 19.9%, which on WMT is that same tie-calibrated accuracy, the tie-blind pairwise accuracy on the selection half 18.7%, the grouped mean 19.5% and the pooled coefficient 23.8%. The reordering is real and it is not small: on WMT the grouped mean errs on 20.9% where acc^*_{eq} errs on 4.7%, the reverse of the ordering under the tie-blind decision. What survives both utilities is that the pooled coefficient is the worst of the coefficients, 23.8% here against 23.6% there.

Candidates from systems of the same quality, with the systems chosen on the selection half. The candidates of a cluster here come from different systems, and an evaluator that recognises the system knows something true about the candidate, which choosing among several samples of one

model does not supply. How much it is worth is measurable: a predictor that outputs only the candidate's system mean label, estimated on the selection half, chooses correctly on 67.4% of the evaluation half's within-cluster pairs and recovers 41.2% of its available best-of-n gain, so this is not a small part of these decisions. `scripts/decision/system_free.py` builds the nearest available condition without touching a label, and it does so split by split: on the selection half the mean label of every system is estimated from that half's clusters alone, the window of half the systems closest together in those means is chosen, 4 to 12 systems, and the window is applied unchanged to the evaluation half, whose labels nothing has read. Choosing the window on the whole benchmark instead leaves the labels untouched but not unused, since the evaluation half's labels then take part in defining the set measured on; that construction is kept beside the present one as a check on how much it matters. The windows are stable, agreeing with the whole-benchmark window at a Jaccard index of 0.76 on average over cells, 0.53 at the least, and identical to it on 37% of splits, with at most 216 distinct windows on any cell. Labels, ties, clusters, evaluators and the selection rules are unchanged, and the rules are still read off the full candidate set on the selection half, which is the practitioner's position. Residualising the labels on system identity would be the obvious alternative and is wrong here: it breaks every tied pair in the direction of the system mean and manufactures an anti-correlation, which on the heavily tied labels drives the same control to *below* chance. The restriction cuts the control, still estimated on the selection half, to 55.6% and 15.6% on the evaluation half, which is the system effect that remains there, and the comparison survives it: over 30 cells of 5 benchmarks the grouped mean errs on 28.2% against 32.9% for pooling, ahead on 23 cells and 5 benchmarks at $p = 0.031$, while against each benchmark's own statistic at 28.9% it is ahead on 2 at $p = 0.812$. Under the whole-benchmark window the same numbers are 29.6% against 34.4% and 30.2%, ahead of pooling on 5 benchmarks, so choosing the window on the selection half moved the conclusion by nothing and the numbers by little. Every rule is more expensive on this harder decision, and FRANK is the case to read with care: its nine systems reduce to four whose spread is small but whose control is still well above chance, and every rule there lands above 46% of pairs, which is a decision with little signal left rather than a comparison of rules.

Coefficient against aggregation. The grouped Spearman mean differs from the pooled Pearson coefficient in two ways at once, the coefficient and the aggregation. Table 6 scores the six cells of that factorial, three aggregations by two coefficients, on the main run's own draws (`scripts/decision/factorial_utility.py`, the pooled Spearman coefficient from the items of the selection set, the within-cluster Spearman as the Pearson coefficient of the within-cluster midranks, the grouped Pearson mean from the per-cluster Pearson table; every cell targets.py also computes is checked against it), by the selection error over evaluator pairs and by the best-of-n recovery of the evaluator each cell selects: at half the clusters the per-cluster mean adds +2.0 points of recovery to the pooled Pearson coefficient and +1.8 to the pooled Spearman coefficient, the within-cluster aggregation +1.2 and +1.7, and the coefficient -0.1 within the pooled aggregation and -0.4 within the grouped one (positive on 12 and negative on 6 configurations for the pooled Spearman against the pooled Pearson coefficient); at eight clusters the per-cluster mean adds +1.7 and +1.0. On the primary configurations the aggregation carries the ordering and the utility alike and the coefficient neither, so the grouped Pearson mean, 48.7% recovery at half against 48.3% for the grouped Spearman mean the paper headlines, is the best cell by utility there; on WMT23 the aggregation loses 0.1 and 1.4 points and on RewardBench 2 at eight prompts the Pearson coefficient is worth 3.0 more than the Spearman coefficient inside the grouped aggregation, and on PRM800K at eight problems every cell but the within-cluster Pearson coefficient loses against pooling, the per-cluster Pearson mean -3.7, the pooled Spearman coefficient -6.4 and the per-cluster Spearman mean -7.3, so there the coefficient and the per-cluster aggregation lose alike over six binary labels a problem: point estimates whose only error is the Monte Carlo error of the split average, without a corpus-bootstrap interval, since the bootstrap streams of the main run carry the five rules and not these cells. Four of the cells are sums over clusters and have since been given the corpus bootstrap of the main run (Table 7): on the primary configurations at half the per-cluster Pearson mean's evaluator is worth +2.0 $[+0.1, +4.0]$ more than the pooled coefficient's and the rank coefficient adds -0.4 $[-1.3, +0.5]$ inside the per-cluster mean, so on these configurations the larger part of what the grouped Spearman mean gains over pooling is the aggregation.

Choosing the rule on five benchmarks and paying on the sixth. `scripts/decision/rule_choice.py` makes the choice the way a reader would: rank the five coefficients on five benchmarks, name the best, and pay for it on the benchmark that took no part.

Table 6: The coefficient crossed with the aggregation, same draws as Table 5: each cell is a selection rule, computed on the selection half or on eight of its clusters, scored by its selection error over evaluator pairs on the evaluation half (Err, %) and by the best-of-n recovery of the evaluator it selects there (Utility, %), point estimates with equal weight over the configurations of the set. Rows differ by the coefficient, columns by the grouping: pooled over items, within clusters after removing the cluster means, and the equal-weight mean of one coefficient per cluster; the last column counts the configurations on which the per-cluster mean's evaluator is worth more / less than that of the pooled cell of the same coefficient. On RewardBench 2, four completions a prompt with one correct and no ties, the within-cluster and per-cluster Spearman cells coincide by construction. The pooled Pearson cell is r_T and the grouped Spearman cell is grp of Table 5; the grouped Spearman mean under the convention that skips a cluster the evaluator ties inside errs on 22.9% and recovers 48.2% on the primary configurations at half. Point estimates; Table 7 carries the corpus-bootstrap intervals of the cells the bootstrap replays.

	Err, %			Utility, best of n, %			grouped >/< pooled
Coefficient	pooled	within	per-cluster mean	pooled	within	per-cluster mean	configurations
primary (19), half the clusters							
Pearson	27.4	25.1	22.9	46.6	47.9	48.7	11/5
Spearman	26.5	22.8	22.3	46.5	48.2	48.3	8/9
primary (19), eight clusters							
Pearson	31.1	28.9	28.4	43.6	45.2	45.3	12/7
Spearman	30.9	28.9	28.8	43.7	44.6	44.7	11/7
WMT23 MQM (3), half the clusters							
Pearson	30.4	20.5	19.2	55.9	56.7	55.8	2/1
Spearman	27.0	18.0	18.0	56.8	54.5	55.3	1/2
WMT23 MQM (3), eight clusters							
Pearson	31.9	29.7	30.2	51.2	51.1	51.5	1/1
Spearman	31.6	30.4	30.8	50.4	51.5	51.1	3/0
RewardBench 2 (5), half the clusters							
Pearson	17.7	17.4	15.8	77.6	78.9	79.7	3/0
Spearman	16.5	15.9	15.9	76.8	79.5	79.5	3/1
RewardBench 2 (5), eight clusters							
Pearson	26.2	27.1	27.8	65.9	66.8	66.5	4/1
Spearman	26.1	30.0	30.0	65.8	63.5	63.5	1/4
PRM800K (2), eight problems							
Pearson	25.3	25.4	29.7	51.0	51.0	47.3	0/2
Spearman	30.1	31.3	32.1	44.6	44.4	43.7	1/1

Table 7: The cells of Table 6 whose statistics sum over clusters, with the corpus-bootstrap interval of Section 4 (200 resampled corpora of 300 splits, the estimate ± 1.96 standard deviations of the mean over the set's configurations): paired differences in points of best-of-n recovery between the evaluators two cells select, on the draws of Table 5, every candidate kept. The first three columns are against the pooled Pearson coefficient r_T; the fourth is what the rank coefficient adds inside the per-cluster mean and the fifth what equal weights per cluster add to the Pearson coefficient. The pooled and within-cluster Spearman cells rank items across the labelled clusters, are not replayed by the bootstrap and keep their point estimates.

Set	Budget	within Pearson − pooled Pearson	per-cluster Pearson mean − pooled Pearson	per-cluster Spearman mean − pooled Pearson	Spearman − Pearson, per-cluster mean	per-cluster mean − within, Pearson
primary (19)	eight	+1.6 [+0.7, +2.6]	+1.7 [+0.6, +2.8]	+1.2 [+0.1, +2.2]	−0.5 [−1.0, −0.0]	+0.0 [−0.4, +0.5]
primary (19)	half	+1.2 [−0.5, +2.9]	+2.0 [+0.1, +4.0]	+1.6 [−0.2, +3.5]	−0.4 [−1.3, +0.5]	+0.8 [−0.2, +1.8]
WMT23 MQM (3)	eight	−0.1 [−1.1, +0.9]	+0.3 [−0.9, +1.5]	−0.0 [−1.4, +1.3]	−0.3 [−1.2, +0.6]	+0.4 [−0.7, +1.5]
WMT23 MQM (3)	half	+0.7 [−1.4, +2.8]	−0.1 [−2.6, +2.3]	−0.6 [−3.6, +2.4]	−0.5 [−2.9, +2.0]	−0.9 [−3.0, +1.3]
RewardBench 2 (5)	eight	+0.9 [−0.2, +2.1]	+0.6 [−0.5, +1.7]	−2.4 [−3.8, −1.0]	−3.0 [−3.9, −2.0]	−0.3 [−1.2, +0.5]
RewardBench 2 (5)	half	+1.3 [−0.1, +2.6]	+2.1 [+0.5, +3.7]	+1.8 [+0.2, +3.5]	−0.3 [−1.1, +0.5]	+0.8 [−0.1, +1.8]
PRM800K (2)	eight	+0.0 [−1.4, +1.5]	−3.7 [−5.6, −1.7]	−7.3 [−9.6, −5.0]	−3.7 [−5.2, −2.1]	−3.7 [−5.4, −2.0]
PRM800K (2)	half	+0.0 [−0.1, +0.1]	−0.1 [−1.7, +1.6]	−0.2 [−2.9, +2.5]	−0.2 [−2.1, +1.8]	−0.1 [−1.7, +1.6]

The decision score measured on the selection half is excluded from what may be named, since it is not a coefficient a benchmark publishes. The grouped mean is named 6 times of six, and the named rule costs 18.8% on the held-out benchmark against 22.8% for that benchmark's own statistic, 28.2% for pooling and 20.4% for the decision score itself, beating the own statistic on 5 of the six. The same script gives the benchmark-level differences unrounded: against each benchmark's own statistic the grouped mean is ahead by FRANK $+5.61$, HANNA $+0.45$, RealSumm $+0.00$, SummEval

+0.03, WMT22 MQM +11.52, speechocean762 +6.23 points, and against the pooled coefficient by FRANK +12.09, HANNA +4.10, RealSumm +19.73, SummEval +0.25, WMT22 MQM +13.73, speechocean762 +6.23. It records the family too: with six benchmarks a one-sided sign test cannot fall below $1/64$, so no comparison here clears a Bonferroni line over the 6 comparisons the family holds, at 0.0083. When the primary comparison was fixed is stated exactly: the grouped mean entered the experiment as a sixth rule after the pooled-against-within results on a hand-chosen subset of seven configurations had been seen (the change log of the release records the order), and its comparison against the pooled coefficient was named the primary one before the criterion-defined set of Appendix H was run, on which the within-cluster coefficient lost its six-of-six and the grouped mean kept it; so the primary comparison is pre-specified relative to the full run and not relative to that subset, which is one reason the paper rests its cross-benchmark claims on effect sizes and directions rather than on the sign test, and everything else in this section is description. Two further readings of the comparison with current practice are in the same script. Leaving WMT out, the own statistics trail the grouped mean by +2.08 points cell by cell over the 32 remaining primary cells and by +2.46 with the five benchmarks weighted equally, against +3.57 with WMT in, of which WMT supplies 51%; the gap to pooling without WMT is +5.15 and +8.48. At the level of the four task families, which is the coarser unit because FRANK, RealSumm and SummEval score CNN/DailyMail summaries and share ROUGE, BERTScore, MoverScore and BARTScore variants, the grouped mean is ahead of pooling on 4 of 4 families and of the own statistics on 4, at family levels of 21.6%, 13.7%, 14.3% and 20.3% against 32.3%, 17.8%, 28.0% and 26.5% for pooling; a sign test over four units cannot fall below $1/16$, so this is description of the dependence rather than a test. One further link is worth recording: across the primary crossed cells the gain from grouping is not concentrated where system identity is most informative, the Spearman correlation between the two being -0.25, and on the half of cells where the system shortcut is smallest the grouped mean saves 7.9 points against 5.3 on the other half.

The held-out unit on WMT: segment, document and domain. The MQM annotation of WMT22 keeps the document: a source document is a run of segments, each segment is rated by one annotator with the whole document in view, and the en-de and en-ru test sets keep the first ten sentences of every document (Freitag et al., 2022). A split that moves whole segments therefore lets two segments of one document, sharing a topic, a difficulty and, as measured below, a rater, sit on both sides. The paper splits by document on the three configurations whose decision is taken inside a source segment, 181, 181, 256 documents for en-de, en-ru and zh-en, and `scripts/decision/doc_blocked.py` measures what the change did, on the same seed (Table 67). It did little: no rule moves by more than 1.1 points on any cell, and over the three configurations the pooled coefficient errs on 28.0% under the document split against 27.7% under the segment split, the grouped mean on 14.3% against 14.0% and the own statistic on 25.8% against 25.8%, so the grouped mean's lead over the own statistic on this benchmark is +11.52 points under the unit that respects the documents against +11.81 under the one that did not. The spread over splits is the second thing a document unit changes, since a spread over segment splits is a claim about fresh segments of the same documents and a spread over document splits a claim about fresh documents; on the pairwise choice the gap between the pooled coefficient and the grouped mean runs from +3.4 to +18.7 points over document splits against +3.3 to +18.2 over segment splits. What a document split leaves shared is measured rather than asserted, since a random document split does not arrange that the two halves share no rater or topic: WMT22 assigned one rater per document on en-de and en-ru and one per document and system on zh-en, seven, four and twelve raters, which give 7, 5, 6 rater blocks on the three pairs once a blank rater forms a block of its own and each zh-en segment takes its modal rater; over the document splits the two halves hold the same rater blocks at a Jaccard overlap of 0.89, 99% of the evaluation half's segments having a rater who also rated segments of the selection half, and the same domains at 1.00, 100%; the documents themselves overlap at 0.00. A split by whole rater blocks, which moves whole documents too on en-de and en-ru, is therefore scored as a third unit, at 500 random halves of the blocks; on zh-en 576 of the segments carry several raters across their translations, so that split is only approximately rater-disjoint: the rules run 27.9%, 20.8%, 15.2%, 25.8% and 12.6%, within -0.11, +0.86 and -0.04 points of the document split for the pooled, grouped and own statistics, and the grouped mean is the cheapest of the four coefficients. Leaving one of the test set's four domains out instead, choosing on three and paying on the fourth, is a different question, a shift of genre and of rater at once with no randomness in it, and it answers differently: the rules run 29.6%, 23.6%, 20.2%, 27.0% and 17.8% for the pooled, within-cluster, grouped, own and decision statistics, and the grouped mean is the cheapest of the four coefficients. Four domains are four folds of one benchmark, which is description;

Table 8: Selection error, percent of evaluator pairs, of eight statistics computed on the selection half and scored under four utilities on the evaluation half, over the 19 primary configurations and the splits of Table 65: the pairwise choice, best of n, best of 3 over every triple, and the three-way choice that also pays for calling two candidates equal. Mean is over the four; a column headed by a minus sign is the mean over the three other payoffs, the same three for every rule; worst is the largest of the four. Underlined: a decision statistic under its own utility. Bold: lowest in the column.

	Paid under				Mean over the other three					
Statistic on the selection half	Pairwise	Best of n	Best of 3	Three-way	Mean	−pair	−best n	−best 3	−3-way	Worst
Coefficients										
r_T, pooled Pearson	19.8	27.3	21.2	23.8	23.0	24.1	21.6	23.6	22.8	27.3
r_W, within-cluster Pearson	14.2	25.1	15.7	18.4	18.4	19.7	16.1	19.2	18.4	25.1
grp, mean per-cluster Spearman	11.9	22.2	14.6	19.5	17.0	18.7	15.3	17.9	16.2	22.2
own, the benchmark's statistic	16.2	25.1	17.4	19.9	19.6	20.8	17.8	20.4	19.6	25.1
Decision statistics measured on the selection half										
pairwise accuracy	**11.1**	22.2	**14.1**	18.7	**16.5**	**18.3**	**14.6**	**17.3**	**15.8**	22.2
best of n	22.5	25.0	22.3	29.0	24.7	25.5	24.6	25.5	23.3	29.0
best of 3	14.2	**22.1**	14.5	21.5	18.1	19.4	16.7	19.3	16.9	**22.1**
acc^*_{eq}, tie-calibrated	17.8	28.9	20.8	**11.3**	19.7	20.3	16.6	19.3	22.5	28.9

what it says is that the document unit is the one this paper's claim needs, that a rater unit changes little, and that a domain shift is a claim it does not make.

The other end of the cluster. Best of n pays for the candidate an evaluator scores highest, and some uses of an evaluator take the lowest instead: which of a learner's sentences to correct first, which summary or translation to flag. `scripts/decision/worst_of_n.py`, written and run after every protocol and graded by none, reruns the engine of Section 4 on the nineteen primary configurations with every label and every evaluator's score negated, on the main run's draws. Negating both leaves the pooled and within-cluster coefficients, the grouped mean, the pairwise accuracy and each benchmark's own statistic unchanged, so on every draw these rules name the evaluators they name in the main run, and their pairwise choice reproduces the main run's estimates exactly; what changes is the payoff, which becomes worst of n, the share of the avoidable drop below a cluster's mean that the candidate an evaluator scores lowest finds, and the decision statistic read off the labelled clusters, which becomes worst of n on them. Under worst of n the grouped mean's pick earns +3.2 $[+0.9, +5.5]$ points over pooling's at half the clusters and +1.2 $[-0.3, +2.6]$ with eight labelled clusters, against +1.6 $[-0.3, +3.5]$ and +1.2 $[-0.1, +2.4]$ under best of n on the same configurations, draws and resamples, and the grouped mean removes +5.0 $[+3.5, +6.4]$ points of pooling's selection error at half against +5.0. On speechocean762, where the decision the introduction names is which sentence to correct, the gain at half is +1.8 $[-1.3, +5.0]$, against +1.9 $[-2.1, +6.0]$ under best of n. The configurations that gain most are not the same ones: across the 19 of them the two gains at half correlate at -0.01 and share their sign on 10, both being zero on two; HANNA relevance, +21.0 points under best of n, gains +1.6 under worst of n, whose largest gain, +22.9, is on HANNA complexity. The same picks, made by the same rules, are worth a different amount depending on which end of the cluster the decision reads. Both payoffs' intervals are corpus bootstraps of 40 splits a replicate on the same resamples, the frozen convention, whose interval for the best-of-n primary mean the calibration puts at 1.47 times the true standard deviation at eight clusters and 1.00 at half (Appendix AA); the main run's, of 300, are narrower at eight. The run's full output is left out of the supplement for its size and regenerated by the script; `worst_of_n_summary.py` writes the summary the paper reads.

Selection statistics scored under another utility. The reference rule of Table 65 reads the decision itself off the selection half. One ground for leaving it out of the rules a benchmark could publish is that it is not a coefficient a benchmark publishes once and several decisions read; a benchmark can publish a pairwise accuracy as easily as a correlation, so the ground has to be measured: does a coefficient serve a decision it was not fitted to better than a decision statistic fitted to a different one does? `scripts/decision/cross_utility.py` answers on the same splits. Eight selection statistics are computed on the selection half, the four coefficients and four decision statistics, the tie-blind pairwise accuracy, best-of-n, best-of-3 over every triple of a cluster, and acc^*_{eq}; four payoffs are measured on the evaluation half, the pairwise choice, best of n, best of 3 and the three-way decision of the tie-aware paragraph at each evaluator's own threshold from the selection half. Over the

19 primary configurations the diagonal is the best rule where the decision statistic is well measured on half the clusters: the pairwise accuracy under the pairwise choice at 11.1% against 11.9% for the grouped mean, and acc^*_{eq} under the three-way decision at 11.3% where the next rule, the within-cluster coefficient, errs on 18.4%. Best-of-n computed on half the clusters is noisy enough that on its own diagonal it errs on 25.0% against 22.1% for best-of-3 and 22.2% for the grouped mean. Off the diagonal the statistics sort by what they are grouped on and not by whether they are decisions. Averaged over the four payoffs, which is what a statistic published once and read by several decisions is asked to do, the pairwise accuracy on the selection half errs on 16.5%, the grouped mean on 17.0%, best-of-3 on 18.1%, the within-cluster coefficient on 18.4% and each benchmark's own statistic on 19.6%, while the tie-calibrated accuracy errs on 19.7%, the pooled coefficient on 23.0% and best-of-n on 24.7%: under the pairwise choice the tie-calibrated accuracy errs on 17.8% and best-of-n on 22.5% against 11.9% for the grouped mean, and under the three-way decision best-of-n errs on 29.0%. The grouped mean is the lowest of the eight on 0 of 6 benchmarks and below the pooled coefficient on 6; its top-1 regret runs 7%, 21%, 12% and 11% of the span under the four payoffs against 16%, 29%, 17% and 21% for pooling. So that ground is half right. A statistic that pays for what the other decisions do not, a tie or the top of a cluster, serves its own utility and no other; a statistic grouped on the decision's cluster, whether a correlation or a pairwise accuracy over every pair inside it, is what a benchmark can publish once and have several decisions read, and among those the pairwise accuracy is not worse than the correlation. Where the utility is known the statistic fitted to it is the rule to compare against, which is how the main text uses it. Table 8 carries the whole matrix with the mean over the four payoffs, the mean over each three of them, the same three for every statistic, and the worst payoff: the lowest mean is the pairwise accuracy on the selection half, the lowest of every three-payoff mean the pairwise accuracy on the selection half, ahead of the grouped mean by at most 0.7 points, and the lowest worst case best-of-3 on the selection half; the four payoffs are weighted equally because nothing says which decision a benchmark's readers will take, and the worst column is the reading that needs no weighting. The coefficient columns of that matrix are the numbers of `selection_rules.py` to 3.3e-16, both scripts drawing the same splits. With all eight statistics as candidates, `scripts/decision/rule_choice.py` repeats the choice of the rule on five benchmarks and the payment on the sixth: under the two primary payoffs the candidates named are pairwise accuracy on the selection half 6, at 18.5% on the held-out benchmark against 18.8% for the grouped mean throughout, 22.8% for the own statistics and 28.2% for pooling; under all four payoffs they are pairwise accuracy on the selection half 6, at 18.2%. Since the ranking of the eight moves with the payoff and the pairwise accuracy and the grouped mean are within a point of each other on every column, the paper does not claim the grouped mean is the statistic to publish; it claims the grouping.

What gets named, and on which pairs. The error rate over every pair of evaluators depends on the panel, and a panel with three ROUGE variants or with weak baselines hands every rule pairs it cannot get wrong. `scripts/decision/winners.py` reports the choice as it is made in practice and stratifies the pairs, on the same splits (Table 66). Named one evaluator each, the rules cost on the 38 primary cells a top-1 regret of 23% of the span between the best evaluator on the evaluation half and the average one for the pooled coefficient, 19% for the within-cluster one, 14% for the grouped mean, 18% for the own statistic and 20% for the decision itself; the grouped mean's regret is below the pooled coefficient's on 5 of 6 benchmarks and below the own statistic's on 4. The pooled coefficient and the grouped mean select different evaluators most often on 20 of the 38 cells, and on those the grouped mean's regret is the lower on 15 and the pooled coefficient's on 5; the grouped mean selects the evaluator that is in fact best on the evaluation half most often on 26 cells, the pooled coefficient on 17 and the own statistic on 22. Restricting the pairs to the competitive ones, both evaluators in the top half of the panel by held-out score, which are 25% of the pairs, the rules err on 31.6%, 28.7%, 26.5% and 29.2% for the pooled, within-cluster, grouped and own statistics, with the grouped mean ahead of pooling on 6 benchmarks; restricting to pairs of different metric families, 90% of the pairs, they err on 23.2%, 19.4%, 16.6% and 20.1%, the grouped mean ahead of pooling on 5; and on pairs that are both, 21% of the pairs, on 32.3%, 29.5%, 26.7% and 29.2%, ahead on 5. A practitioner who first shortlists the top half of the panel by the decision measured on the selection half and then chooses among the shortlist by a published statistic pays 41% of the shortlist's span with the pooled coefficient and 27% with the grouped mean, the latter below the former on 5 benchmarks. The families are assigned by name and printed with the results so that the assignment can be disputed.

Does the share predict the cost, with everything else fixed? On the real benchmarks the contrast between cells above and below a label share of one half separates benchmarks rather than labels, because every configuration of a benchmark falls on one side, so the share and the benchmark's identity are confounded there. `scripts/sim/screen_simulation.py` removes the confound in the two ways available. In simulation, one draw fixes the within-cluster part of the labels, the cluster effects and a panel of twelve evaluators whose between-cluster validity, within-cluster validity and between-to-within scale are drawn independently, and the share is then set by the mixing weight alone, so the within-cluster ordering of the items, and with it the decision, is the same at every share. Over 200 such benchmarks at each of 9 shares, the cost of pooling, the pooled coefficient's held-out error minus the grouped mean's, rises with the share at every one of the 8 steps, from 8.9 points at a share of 0.11 to 32.6 at 0.91, and is positive on at least 95% of the benchmarks at every share. Two things the screen was asked to do it does not: inside one share the in-sample share on the selection half predicts the cost across benchmarks at a Spearman correlation of only +0.05 to +0.16, which is sampling noise predicting sampling noise, and a threshold chosen on half the benchmarks to separate expensive from cheap lands at 0.80 rather than at one half, with costs of 31.3 against 17.8 points on the other half either side of it and a Spearman correlation of +0.69 there. There is no line: the cost is a slope in the share, positive from the lowest share up. The simulation draws each evaluator's between-cluster validity and within-cluster validity independently, which limits what it can say about panels where the two go together; drawn instead from a Gaussian copula with correlation -0.8, -0.4, 0, 0.4 and 0.8 and the same marginals, over 200 panels at shares of 0.3, 0.6 and 0.9, the cost of pooling runs from 3.6 to 53.9 points over the fifteen cells, smaller where the correlation is positive, since a pooled coefficient that reads the between-cluster validity then learns something true about the within-cluster one, and larger where it is negative, and it stays positive in every cell, on at least 78% of the panels; at a share of 0.9 it is 13.9 points at $+0.8$ and 53.9 at -0.8. The same manipulation is available on the real labels: with $\bar{y}_c$ the cluster means and $\bar{y}$ the grand mean, the labels $\bar{y} + \lambda(\bar{y}_c - \bar{y}) + (y_i - \bar{y}_c)$ keep every within-cluster difference, so every decision, r_W and every grouped statistic of every evaluator are unchanged, and only the between-cluster part moves, the share becoming $\lambda^2\mathrm{SSB}/(\lambda^2\mathrm{SSB} + \mathrm{SSW})$. `scripts/sim/share_rescale.py` runs the held-out experiment on the transformed labels of every primary configuration at $\lambda \in \{0, 0.25, 0.5, 1, 2, 4\}$, on the same splits and with the evaluators' outputs untouched, and checks the invariance numerically: the within-cluster and grouped errors and every decision score on every evaluation half agree across λ to 2.3e-04. The share runs from 0.00 to 0.81 on average and the cost of pooling from 6.5 to 10.2 points, the pooled error from 23.5% to 27.2% against a grouped error fixed at 17.1%; the cost does not vanish at $\lambda = 0$, where the labels have no between-cluster part, because the pooled coefficient is then the within-cluster correlation scaled by each evaluator's own between-cluster share, which still ranks evaluators by something the decision does not read; the cost rises on 72 of the 95 steps, is monotone on 9 of the 19 configurations, is positive on at least 17 configurations at every λ, and is higher at the largest λ than at the smallest on 5 of 6 benchmarks. This is a controlled transformation of real scores and not a natural change of distribution, and it is the evidence that the share moves the cost with everything else fixed on these benchmarks. Inside each real benchmark a coarser contrast is also reported, an association and not a manipulation: the clusters of every primary configuration are ordered by their mean label, the inner half, closest to the median, has a small between-cluster share and the outer half a large one, at 0.08 against 0.44 on average, and the held-out experiment is run inside each half. The two halves are different sets of clusters, which can differ in difficulty, in the spread of their labels and in which candidates they hold as well as in their share, so this is an association. The outer half costs pooling more on 11 of the 19 configurations and 5 of 6 benchmarks, at $p = 0.109$, 5.7 points against 3.3, the share and the cost correlating at +0.40 over the 32 halves and their within-configuration differences at +0.20; and the cost is positive on 17 of the 19 inner halves, whose shares reach at most 0.26. What this supports is narrower than a recommendation: the share says how much a within-cluster decision cannot read off the pooled number, and in these manipulated worlds pooling's selection error over evaluator pairs is the larger at every share the two manipulations produce; whether the evaluator it selects is worth less is the question of Sections 5 and 7, and the reporting protocol of Appendix D rests on that section and not on this one.

J Three Measurement Targets

The target, and the compression an independent measurement set applies. Take the clusters of a configuration as drawn from a population, and for a pair of evaluators let d_c be the difference

of their per-cluster payoff contributions on cluster c: for the pairwise choice the concordant pairs of c, for best of n the gain recovered there. The payoff on a set S of clusters is a ratio of sums over S, and for two evaluators compared on the same S the denominator is shared, so the pair's payoff difference on S has the sign of $D_S = \sum_{c \in S} d_c$, and the better evaluator in the population is the sign of $\delta = \mathbb{E}[d_c]$. A rule reads a set S of b clusters and names an evaluator; its deployment error on the pair is $e = \Pr[\text{the rule selects the evaluator with the sign opposite to } \delta]$, over draws of S. Now pay the rule by a measurement set M independent of S, and let $q = \Pr[\operatorname{sgn} D_M \neq \operatorname{sgn} \delta]$ be M's own error on the pair. The rule is scored wrong when it agrees with the population and M does not, or the reverse, so

$$\Pr[\text{scored wrong against } M] = e(1-q) + (1-e)q = e + q(1-2e), \tag{2}$$

a pull toward one half from either side: a rule with $e < \frac{1}{2}$ is charged $q(1-2e)$, most when it is most accurate, and one with $e > \frac{1}{2}$, a statistic that orders the pair against the payoff systematically, is credited. Three conditions are part of the statement. It is conditional on M being decisive on the pair: the error is scored only on pairs whose payoffs on M differ, a pair tied on M is dropped for every rule, and q is the probability of the wrong winner given that M is decisive, which on binary labels and a small M is not a formality. It needs $q < \frac{1}{2}$, which a positive mean difference δ does not guarantee on a finite set: with $d_c = -1$ with probability 0.9 and $+10$ with probability 0.1, $\delta = 0.1$, while a set of five clusters has a negative sum with probability $0.9^5 = 0.59$. And it is a statement about one pair. For two rules A and B on the same pair the observed difference is $(1-2q)$ times the true one, so their order on that pair is kept when $q < \frac{1}{2}$; averaged over the K pairs of a panel,

$$\bar{e}_A^{\,\mathrm{obs}} - \bar{e}_B^{\,\mathrm{obs}} = \frac{1}{K} \sum_{j=1}^{K} (1-2q_j)\,(e_{A,j} - e_{B,j}), \tag{3}$$

a reweighting of the true gap by how reliably M decides each pair, which can reverse the average order even when every $q_j < \frac{1}{2}$. Take three evaluators with $q = 0.49$ on the pair of the first two and $q = 0$ on the other two pairs, rule A wrong on the first pair with probability 0.3 and rule B wrong on the third with probability 0.2: the true mean errors are 10% and 6.7%, so B is better, and the observed ones 16.5% and 23%. Nor is a common q_j enough: only the pairs decisive on M are averaged, so the set of pairs, and with it the weight $1/K_M$, is random as well, and the mean can reverse with every $q_j = 0$. Three evaluators with population payoffs $(0,\ 0.1,\ 1)$, a measurement set that with probability 0.9 reads $(0, 0, 1)$, tying the first two, and with probability 0.1 reads $(0, 1, 1)$, tying the last two, so that every decisive pair is ordered as the population is and every $q_j = 0$; rule A reverses only the first pair with probability 0.8 and rule B only the last with probability 0.7, independently of M: the true mean errors over the three pairs are 26.7% and 23.3%, so B is better, and the observed ones 4.0% and 31.5%, because the pair each rule gets wrong is the pair the measurement set mostly drops. The observed mean gap is $\sum_j (e_{A,j} - e_{B,j})\, \mathbb{E}_M[I_j(M) s_j(M)/K_M]$ over the sets with $K_M > 0$, I_j the indicator that M decides pair j and s_j the sign of its decision against the population, and neither the common q_j nor a fixed K_M makes it a positive multiple of the true gap. The paper's primary quantity is therefore the utility of the selected evaluator, a ratio on each split whose gap to a linear reading is measured below, and selection error is read as an attenuated and reweighted explanation. How often the reweighting reversed an average order in the simulation below is counted there: 12 of 640 comparisons of two rules against the fixed set, 10 against a fresh sample and 18 against the complement, all with a true gap of at most 0.1 points, against 30 against the in-sample target, with true gaps up to 8.5 points.

The estimand on the paper's own splits. Two checks, `scripts/decision/estimand_checks.py` on the same document-blocked half-splits as the main run (Table 9). The random denominator: the payoff on M is $\sum_{c \in M} g_c / \sum_{c \in M} w_c$, and for every evaluator pair the mean over 300 splits of the pair's difference on M is set against the pair's difference on the whole corpus, in points; the same for the unnormalised per-cluster mean gain $\sum_{c \in M} g_c / |M|$, which is linear in M and whose gap is sampling noise only, the floor against which the ratio's gap is read. Over the 2267 pairs of the primary configurations the ratio's gap exceeds twice that floor on 4 of 19 configurations and reaches 5.7 points on SummEval consistency, a label under which most clusters hold almost no recoverable gain and a few hold most of it, the case of the reader's two-cluster example; the mean over splits and the whole-corpus difference disagree in sign on 13 of the 2267 pairs under best of n and 10 under the pairwise choice, whose largest gap is 0.4 points. Paid in a

Table 9: The estimand checks on the main run's own 300 half-splits and naming priorities, replayed from its streams so that the ratio column below is the grp$-r_T$ column of Table 5 to the last digit, every primary and validation configuration. gap: the largest, over evaluator pairs, of the absolute difference between the mean over splits of a pair's recovery difference on the evaluation half and the pair's difference on the whole corpus, in points; floor: the same for the unnormalised per-cluster mean gain, which is linear in the evaluation half and differs from its corpus value by sampling noise only; flips: pairs on which the two differences disagree in sign; grp$-r_T$: the paired gain of the grouped mean's evaluator over pooling's, paid as recovery (ratio) and as the per-cluster mean gain over the corpus's mean recoverable gain (linear); dec.: the lowest, over pairs the whole corpus decides, of the share of splits on which the evaluation half decides the pair; gap in error of the grouped mean against pooling averaged over the pairs decisive on the evaluation half (on M) and over the corpus-decisive pairs with a tie charged half an error (fixed). Primary configurations: gap above twice the floor on 4 of 19, headline +1.6 (ratio) against +1.6 (linear), order under the fixed pair set kept on 19 of 19.

			random denominator			grp$-r_T$, points			grp$-r_T$ error	
Configuration	C	pairs	gap	floor	flips	ratio	linear	dec. %	on M	fixed
so762 accuracy	125	36	0.8	1.0	0	+0.0	+0.0	91	+0.3	+0.3
so762 fluency	125	36	1.5	1.4	1	+2.9	+2.8	81	−4.8	−3.7
so762 prosodic	125	36	1.0	1.2	1	+4.7	+4.6	84	−8.6	−9.2
so762	125	36	2.0	2.0	0	+0.0	+0.0	87	−1.9	−1.9
FRANK factuality	499	55	1.1	0.2	1	+6.4	+6.5	99	−11.2	−11.2
HANNA coherence	96	171	1.5	1.4	1	+0.6	+0.6	95	−0.1	−0.1
HANNA complexity	96	171	0.6	0.7	1	+4.8	+4.8	95	−4.7	−4.7
HANNA empathy	96	171	2.0	1.9	1	−2.5	−2.5	93	+1.7	+1.7
HANNA engagement	96	171	0.9	0.9	1	−2.3	−2.3	93	−0.1	−0.1
HANNA relevance	96	171	1.2	1.5	1	+21.0	+21.0	94	−10.3	−10.5
HANNA surprise	96	171	1.1	1.1	0	+0.3	+0.3	95	+0.8	+0.8
RealSumm	100	120	0.7	0.2	0	+2.4	+2.4	96	−16.1	−16.1
SummEval coherence	100	120	0.8	1.1	0	−7.3	−7.3	96	+0.5	+0.5
SummEval consistency	100	120	5.7	1.8	1	+4.8	+4.7	80	−0.7	−0.7
SummEval fluency	100	120	2.1	0.6	1	−6.3	−6.3	81	+0.1	+0.1
SummEval relevance	100	120	1.4	1.4	2	+0.0	+0.0	95	+6.8	+6.8
WMT22 ende	1315	153	0.3	0.4	0	−0.0	+0.0	100	−14.2	−14.2
WMT22 enru	1315	136	0.5	1.4	1	+1.0	+1.0	100	−14.3	−14.3
WMT22 zhen	1875	153	0.8	2.2	0	+0.2	+0.2	100	−18.7	−18.7
WMT23 ende	460	496	0.8	5.4	2	−3.2	−3.2	100	−12.3	−12.3
WMT23 heen	820	496	0.7	1.7	1	−3.3	−3.3	100	−16.1	−16.1
WMT23 zhen	1177	496	0.5	2.0	1	+4.7	+4.6	100	−9.0	−9.0
RB2 factuality	475	15051	0.9	0.6	26	+0.8	+0.8	84	−3.1	−3.2
RB2 precise if	160	15051	1.7	1.2	60	+0.0	+0.0	72	−3.0	−3.2
RB2 math	183	15051	1.0	0.7	10	+7.5	+7.5	47	+1.8	+1.3
RB2 safety	450	15051	0.8	0.6	3	+1.5	+1.5	62	−2.4	−2.5
RB2 focus	495	15051	0.8	0.6	16	−0.6	−0.6	78	−2.1	−2.2

linear unit, the per-cluster mean gain of the selected evaluator over the corpus's mean recoverable gain, which is a constant, the grouped mean's evaluator earns +1.6 points more than pooling's over the primary configurations against +1.6 under the ratio, the same sign on 18 of 19. The pair set: the selection error is re-averaged over the pairs the whole corpus decides, a pair tied on M charged half an error to every rule, so that which pairs enter no longer depends on M; against the pairs decisive on M, the convention of the main text, the mean error of the pooled coefficient, the within-cluster coefficient and the grouped mean over the primary configurations moves from 27.4, 25.1 and 22.3% to 27.3, 25.1 and 22.3%, the grouped mean's gap to pooling keeps its sign on 19 of 19 configurations and changes by at most 1.1 points, and every corpus-decisive pair is decisive on at least 80% of the splits. The two examples above are therefore exact and, on these panels, idle.

The complement. Under the complement target the evaluation side is the other half $\bar{S}$ of a half-split and the rule reads b clusters of the selection half. Over corpora, the selection half and the other half are two disjoint samples of one population, so $\bar{S}$ is a measurement set independent of the reading and Equation 2 applies with q the error of a sample of $C/2$ clusters. On any one corpus the two halves are not independent: $D_S + D_{\bar{S}}$ is the constant D_{all} over the corpus's C clusters, and with the C values d_c treated as a finite population of variance σ^2 and S a uniform subset of b clusters inside a uniform half, $\mathrm{Cov}(D_S, D_{\bar{S}}) = -b\,\frac{C}{2}\,\sigma^2/(C-1)$, $\mathrm{Var}(D_S) = b(C-b)\sigma^2/(C-1)$ and $\mathrm{Var}(D_{\bar{S}}) = \frac{C^2}{4}\sigma^2/(C-1)$, so the two sides correlate at $-\sqrt{b/(C-b)}$, which is -1 at $b = C/2$; `scripts/sim/split_correlation.py` measures -0.29 at eight clusters of a hundred against -0.29 from the display and -1.00 at half. A pair whose D_{all} is near zero is therefore ordered oppositely by the two sides however the labels fall, and the statistic that is the payoff is scored wrong on it with probability near one at half the clusters, while on a pair the corpus

separates clearly it is scored right with probability near one; the smallest example is four clusters with $d = (1,\ 0.8,\ -0.9,\ -0.7)$, where the decision statistic on one half disagrees with the other half on four of the six splits. The two effects offset in expectation, which is what independence over corpora says: on one corpus the complement concentrates the compression of Equation 2 on the pairs that corpus leaves near-tied rather than spreading it at random, and a single-corpus reading of a rule that tracks the payoff closely can sit above or below the fresh-set reading by the luck of which pairs the corpus leaves near-tied. The complement's number is therefore neither the deployment error itself nor a systematic charge on the decision statistic.

The in-sample target. Paying the rule read off $S \subset$ corpus by the decision over the whole corpus gives a target that is not independent of the reading. The corpus's own winner, $\operatorname{sgn} D_{\text{all}}$, errs against the population with some probability, and the reading of a statistic that is the payoff, D_S, shares that error through the b clusters it holds, most at $b = C$, where the statistic is the target and the error is zero by construction; a statistic whose per-cluster contributions track d_c less closely shares less of it and is charged for the rest. So the in-sample target credits the payoff's own statistic and charges the coefficients, and the credit grows with the budget. It is the only target on which the limit error and the flip rate of Section 4 are defined, and they are read from it; error rates are not.

Simulation with a known population. `scripts/sim/targets_sim.py` draws 60 panels of 8 evaluators whose between- and within-cluster validities are drawn independently on $[0.3, 0.9]$ with a between-to-within scale drawn log-uniformly on $[\frac{1}{2}, 2]$, labels with equal between- and within-cluster variance over clusters of 8 items, and for each panel a corpus of $C \in \{32, 100, 400\}$ clusters, a fixed measurement set of $2C$ clusters and a population of 5000, all from one process. Over 60 half-splits of every corpus, the same reading of b clusters of the selection half is paid five ways: against the corpus (in-sample), against the other half (complement), against a fresh sample of $C/2$ clusters drawn from the population at every split, against the fixed set, and against the population (true). Table 10 gives a hundred clusters with half labelled under best of n. The fixed set reads every rule within 0.00 points of what Equation 2 predicts pair by pair from the true errors and its own error $q = 5.7\%$, the fresh sample within 0.1 of the prediction with $q = 11.0\%$, and the complement within 0.6 points of the fresh sample of its size, which is the independence over corpora above, while the pair-by-pair prediction, which takes the complement's winner as independent of the reading on each corpus, misses the complement by up to 4.5 points, which is the anti-correlation; all three keep the rules' true order, the within-cluster coefficient first and the pooled coefficient last, except between the grouped mean and the pairwise accuracy, which are within a tenth of a point of each other in truth, and read the grouped mean ahead of best of n on the labelled clusters by -3.0 points on the fixed set, -2.3 on the fresh sample and -2.3 on the complement against -4.2 in truth. The in-sample target reads best of n on the labelled clusters at 7.8% against a true 11.1%, a credit of -3.3 points, and the grouped mean at 9.2% against 6.9%, a charge of +2.2, so it puts the two +1.4 points apart with the sign the truth does not have, best of n on the labelled clusters first of the five rules where the truth has it fourth; the pooled coefficient it reads at 29.5% against a true 27.8%. At eight labelled clusters the in-sample credit to best of n on the labelled clusters is -1.3 points and at 32 clusters -2.6; at 32 and 400 clusters with half labelled it is -4.9 and -1.5. Under the pairwise choice at a hundred clusters and half labelled the true errors are 27.7%, 6.9% and 6.8% for pooling, the grouped mean and the pairwise accuracy on the labelled clusters, the fixed set reads 28.4%, 8.1% and 8.1%, and the in-sample target 28.2%, 5.2% and 5.0%. The evaluator each rule selects is the same whichever target pays it, and its true regret at a hundred clusters and half labelled is 28%, 1%, 2% and 6% of the population span for the pooled and within-cluster coefficients, the grouped mean and best of n on the labelled clusters. On this panel, whose evaluators differ in one within-cluster validity each and whose grouped mean therefore has almost no limit error under best of n, the grouped mean is the better rule at half the clusters in truth; the released panels, with three ROUGE variants or two checkpoints of one system, hold far more near-tied pairs than eight evaluators drawn over $[0.3, 0.9]$, which is why the compression there is larger.

The targets on the real configurations. `scripts/decision/targets.py` draws 300 half-splits of every configuration, whole clusters and whole documents where clusters nest in them, and inside the selection half nested subsets of b distinct clusters at every budget of the grid; every rule is read off the subset once and paid against the corpus and against the other half, and on PPE against the measurement set, so the three readings differ in nothing but the target; the statistics on a subset

Table 10: Five ways to pay the same reading, in simulation: 60 panels of 8 evaluators, corpora of a hundred clusters with half labelled, 60 splits each, a population of 5000 clusters. Error over evaluator pairs, in percent, of each rule computed on the selection half: true, against the population; fixed, against a measurement set of 200 clusters fixed once per corpus, beside what Equation 2 predicts from the true error and the set's own error q pair by pair; fresh, against a sample of 50 clusters drawn from the population at every split; complement, against the corpus's other half; corpus, against the decision over all hundred clusters, the read ones included. Regret: the true regret share of the evaluator the rule selects, which does not depend on the target.

		Fixed set					
Rule	True	read	predicted	Fresh	Complement	Corpus	Regret
Best of n; the target's own error q: fixed set 5.7, fresh sample 11.0, complement 11.1, corpus 8.0							
pooled r_T	27.8	27.6	27.6	29.4	30.0	29.5	28
within-cluster r_W	5.3	7.5	7.5	12.2	12.3	8.1	1
grouped mean	6.9	9.0	9.0	13.0	13.1	9.2	2
pairwise accuracy on the labelled clusters	6.9	9.0	9.0	12.9	13.1	9.1	2
best of n on the labelled clusters	11.1	12.0	12.0	15.3	15.4	7.8	6
Pairwise choice; the target's own error q: fixed set 3.9, fresh sample 7.0, complement 7.0, corpus 4.7							
pooled r_T	27.7	28.4	28.4	28.4	28.8	28.2	31
within-cluster r_W	5.1	6.6	6.6	8.6	8.8	5.1	2
grouped mean	6.9	8.1	8.1	9.9	10.1	5.2	2
pairwise accuracy on the labelled clusters	6.8	8.1	8.1	9.8	10.0	5.0	2
best of n on the labelled clusters	11.1	11.4	11.5	12.8	13.1	10.8	7

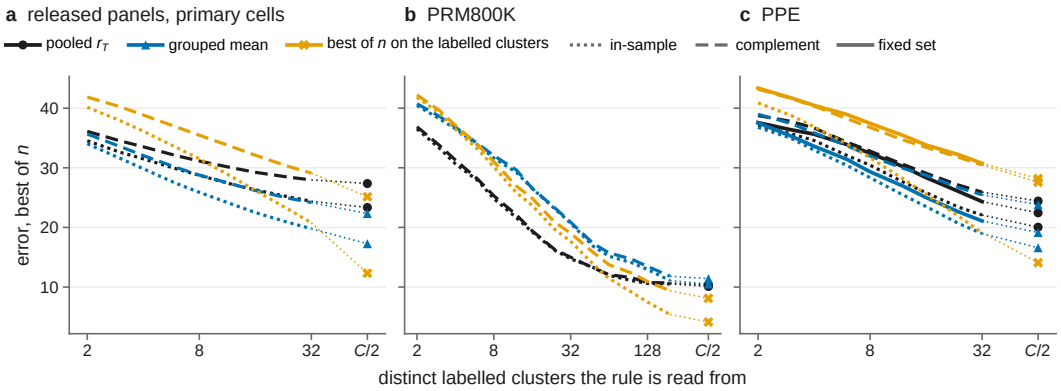


Figure 5: **The three targets on the same draws.** Selection error under best of n of best of n computed on the labelled clusters, the grouped mean and the pooled coefficient, against the budget, paid by the in-sample target (dotted), the complement (dashed) and, on PPE, the fixed measurement set of 192 disjoint prompts (solid); cell-equal means over the configurations of each group, the last point at half the clusters.

are the weighted sums of the per-cluster parts, checked against the mask functions of the held-out script on random masks. On the primary configurations at half the clusters the in-sample target reads the pooled coefficient, the within-cluster coefficient, the grouped mean and the pairwise accuracy at 18.0%, 11.1%, 6.7% and 5.6% under the pairwise choice and the complement at 19.8%, 14.1%, 11.8% and 11.1%; under best of n the in-sample target reads the pooled coefficient, the grouped mean and best of n on the labelled clusters at 23.3%, 17.3% and 12.3% and the complement at 27.4%, 22.3% and 25.2%, the complement above the corpus by +4.0, +5.1 and +12.8 points. On PPE at 48 prompts under best of n the three targets read best of n on the pool at 14.1%, 27.6% and 28.2% and the grouped mean at 16.6%, 23.8% and 19.1%; the measurement set of 192 prompts is the least noisy of the three evaluation sides and reads the largest gap. Paid by the in-sample target with b clusters drawn with replacement (`scripts/decision/population_protocol.py`, kept in the release), best of n on the labelled clusters reads level with the grouped mean at half the clusters; that reading is the credit measured above, and a budget of 48 draws with replacement holds 37.9 distinct clusters of 96 on average, which the present draws, without replacement, do not.

Agreeing components, and a counterexample built as data. Agreement of the two components is not sufficient, because η_p weights them: the simulation script builds twenty clusters of six items with $\eta_y^2 = \frac{1}{2}$ exactly and two evaluators in the orthogonal between and within subspaces, one with $r_B = 1.00$, $r_W = 1.00$ and $\eta_p = 0.99$, which reaches $r_T = 0.800$, and one with $r_B = 0.90$, $r_W = 0.90$ and $\eta_p = 0.71$, which reaches $r_T = 0.900$; the first orders 100% of the within-cluster pairs correctly and recovers 100% of the best-of-n gain against 86% and 94% for the second, and the pooled coefficient prefers the second. The disagreement share and the limit error of the pooled coefficient in Table 11 are read off the pooled coefficient's ordering against the within-cluster one and against the payoff, and so count this case.

The smallest budgets, audited. The conventions are the same for every rule and every target: every rule is scored on the same decisive pairs of a target at every budget, an undefined statistic chooses at random and costs half an error, and a rule with no defined statistic names an evaluator at random for the top-1 payoff. How often that happens is recorded: at two clusters, on the primary configurations, the pooled coefficient is undefined on 1.8% of scored pairs and the within-cluster coefficient on 2.1%, the grouped mean and the decision statistics on 1.2%, and no rule is undefined on more than 2.1% of pairs at any common budget, below half a percent from 4 clusters on; at half the clusters no rule is undefined on more than 1.14% of pairs, which is why the held-out experiment of Appendix H, which drops a pair on which a rule is undefined instead of scoring it a coin, agrees with the complement target of `targets.py` at half the clusters to a tenth of a point on the primary means. On PRM800K, where 37% of problems carry six equal answer labels, the within-cluster coefficient is undefined on 6.5% of pairs at two problems and below half a percent from 6; on PPE, below half a percent from 2 prompts. Scoring every rule on the pairs on which all five are defined instead moves no aggregate error by more than 0.3 points on the primary configurations, 1.2 on PRM800K and 0.1 on PPE (`scripts/decision/label_budget.py`, which keeps the audit). A budget of b is b distinct clusters, since a cluster is what a benchmark labels and what the choice turns on, and b clusters are not the same labelling cost across benchmarks: 462 items at 32 clusters on average over the primary cells, and on PPE sixteen labelled responses a prompt, so 48 prompts are 768 labels.

K THE LABEL BUDGET

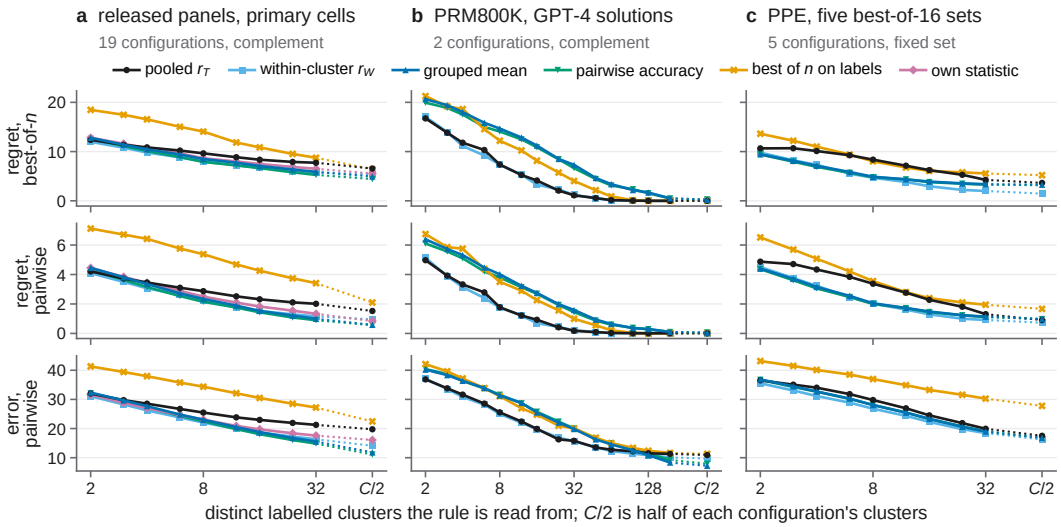


Figure 6: **The label budget, scored out of sample.** Every rule is computed on b distinct clusters of the selection half, 300 splits (PPE 500 draws), and the evaluator it selects is scored on the other half (every set but PPE) or on the fixed evaluation set of 192 prompts a task (PPE). Rows: regret of the selected evaluator under both payoffs, and selection error over evaluator pairs; exploratory configurations weigh equally (Appendix K).

Table 11 gives every configuration: the budget grid is $b \in \{2, 3, 4, 6, 8, 12, 16, 24, 32, 48, \dots\}$ up to half the clusters and then every cluster, 300 splits at each, and the five rules are the pooled and

Table 11: The label budget by configuration. Unit: what the half-split moves, whole clusters or whole documents (WMT segments nest in documents), the budget being distinct clusters of the selection half. Limit: the pooled coefficient's error when it reads every cluster, in percent, under the pairwise choice and under best of n, the share of pairs it orders against the payoff on the whole corpus. From: the smallest budget of labelled clusters from which one rule errs less than another at that budget and at every larger one up to half the clusters, against the out-of-sample target, the complement (PPE: the fixed evaluation set): grp $< r_T$, the grouped mean against the pooled coefficient, and pair $< r_W$, the pairwise accuracy on the labelled clusters against the within-cluster coefficient, under the pairwise choice; grp $< r_T$ and own $<$ grp, best of n on the labelled clusters against the grouped mean, under best of n; a dash where the first never stays below the second. At half: the error of r_T / grp / the payoff's own statistic at half the clusters, in percent, against the out-of-sample target (out) and against the in-sample one (in). 300 splits per configuration (PPE 500 draws).

				Limit of r_T		Pairwise: from		Best of n: from		Pairwise at half		Best of n at half	
Configuration	C	Ev.	Unit	pair	bon	grp $< r_T$	pair $< r_W$	grp $< r_T$	own $<$ grp	out	in	out	in
speechocean762 accuracy	125	9	cluster	19	23	2	–	–	–	20 / 18 / 18	20 / 9 / 9	22 / 22 / 23	17 / 17 / 11
speechocean762 fluency	125	9	cluster	22	24	8	8	2	–	22 / 8 / 8	21 / 4 / 4	43 / 38 / 66	26 / 37 / 31
speechocean762 prosody	125	9	cluster	19	26	12	12	2	–	22 / 11 / 12	18 / 6 / 6	37 / 29 / 46	26 / 13 / 21
speechocean762 total	125	9	cluster	17	20	6	16	2	–	20 / 16 / 15	15 / 8 / 7	25 / 23 / 27	19 / 14 / 13
FRANK factuality	499	11	cluster	27	47	24	192	2	–	30 / 16 / 15	28 / 9 / 8	47 / 36 / 40	49 / 31 / 20
HANNA coherence	96	19	cluster	10	13	2	–	2	–	15 / 12 / 11	12 / 7 / 5	17 / 16 / 17	13 / 12 / 8
HANNA complexity	96	19	cluster	10	13	2	12	2	–	14 / 7 / 7	12 / 4 / 3	16 / 11 / 14	14 / 7 / 7
HANNA empathy	96	19	cluster	9	9	2	–	–	–	13 / 10 / 9	11 / 6 / 5	14 / 16 / 17	9 / 12 / 8
HANNA engagement	96	19	cluster	15	14	2	–	2	half	16 / 11 / 10	16 / 6 / 5	18 / 17 / 17	15 / 14 / 9
HANNA relevance	96	19	cluster	15	22	2	8	2	–	23 / 12 / 12	20 / 7 / 6	27 / 16 / 20	23 / 10 / 9
HANNA surprise	96	19	cluster	12	16	2	–	–	–	19 / 12 / 12	16 / 6 / 6	22 / 23 / 24	18 / 17 / 12
RealSumm LitePyramid	100	16	cluster	36	44	2	8	2	–	38 / 14 / 14	37 / 7 / 7	44 / 28 / 40	42 / 18 / 20
SummEval coherence	100	16	cluster	9	18	24	48	–	24	11 / 10 / 9	9 / 6 / 5	17 / 18 / 15	17 / 16 / 8
SummEval consistency	100	16	cluster	11	15	6	3	8	–	13 / 8 / 7	12 / 5 / 4	22 / 21 / 22	16 / 16 / 11
SummEval fluency	100	16	cluster	12	23	6	6	–	24	14 / 12 / 11	13 / 7 / 6	25 / 25 / 21	24 / 23 / 10
SummEval relevance	100	16	cluster	18	13	2	–	–	24	22 / 20 / 20	20 / 10 / 10	21 / 27 / 23	15 / 20 / 11
WMT22 en-de	1315	18	doc.	16	37	192	16	3	192	17 / 14 / 8	16 / 12 / 4	37 / 23 / 15	36 / 21 / 8
WMT22 en-ru	1315	17	doc.	18	32	3	2	3	256	18 / 6 / 5	18 / 4 / 2	33 / 18 / 16	32 / 16 / 8
WMT22 zh-en	1875	18	doc.	27	31	8	48	6	half	28 / 7 / 7	28 / 4 / 3	35 / 16 / 16	33 / 14 / 8
PRM800K, answer label	458	5	cluster	10	0	32	48	–	96	13 / 6 / 7	14 / 2 / 3	8 / 10 / 7	5 / 8 / 4
PRM800K, human step label	457	5	cluster	10	20	half	–	half	2	9 / 9 / 9	6 / 5 / 5	13 / 13 / 9	15 / 13 / 4
PPE MMLU-Pro	96	6	prompt	20	20	–	–	–	16	10 / 18 / 18	13 / 14 / 12	17 / 25 / 23	13 / 14 / 11
PPE MATH	96	6	prompt	0	14	–	–	–	–	6 / 7 / 7	6 / 7 / 7	6 / 7 / 21	18 / 17 / 15
PPE GPQA	96	6	prompt	7	27	–	2	–	–	33 / 35 / 37	14 / 14 / 14	33 / 35 / 53	24 / 24 / 20
PPE IFEval	96	6	prompt	7	7	2	12	2	–	24 / 21 / 20	17 / 7 / 8	29 / 14 / 25	17 / 7 / 12
PPE MBPP-Plus	96	6	prompt	13	33	2	3	2	–	14 / 3 / 3	14 / 3 / 3	28 / 15 / 19	28 / 21 / 13

within-cluster coefficients, the grouped mean, and the pairwise accuracy and best-of-n recovery inside the labelled clusters; each is scored under both payoffs against the complement (on PPE, the fixed measurement set) and against the corpus by the error over evaluator pairs, the flip rate, the limit error, whether the one evaluator it names is the best of the panel, that evaluator's regret in the payoff's units and as a share of the span between the best and the mean evaluator, and the payoff it realises. Figure 6 averages configurations with equal weight over the budgets every configuration holds, up to 32 clusters on the primary cells, the smallest half being 48 clusters. The sampling block is the unit the split moves: on the 3 WMT configurations the decision cluster is the source segment and the block its document, 181 to 256 documents a configuration, so the b segments of a budget come from the selection half's documents, 8 documents on average at eight segments and 26 at 32, and the evaluation half holds no document the rule read; elsewhere the block is the cluster. Against the complement, under the pairwise choice over the 9 common budgets, the within-cluster coefficient is the best of the five on 5 budgets, the pairwise accuracy on 4 and the pooled coefficient on 0, and best of n on the labelled clusters is the worst on 9; under best of n the within-cluster coefficient is the best on 3 and the pairwise accuracy on 6, best of n on the labelled clusters the worst on 9 and the pooled coefficient on 0. Read off every cluster the five err against the corpus on 17.1%, 9.8%, 2.7%, 0.0% and 17.0% under the pairwise choice and on 23.3%, 20.3%, 15.9%, 16.7% and 0.0% under best of n, which are the limit errors. The pooled coefficient's limit error under the pairwise choice runs from 9.2% to 35.8% over the primary configurations and is zero on 0 of them; its Spearman correlation over configurations with the budget from which the grouped mean stays ahead of pooling against the complement is +0.23, and with the disagreement share +0.65; against the complement the grouped mean never stays ahead of pooling on 0 configurations under the pairwise choice and 6 under best of n, and the payoff's own statistic never stays ahead of the grouped mean on 5 and 12, against 2 under best of n against the corpus, which is the credit of Appendix J configuration by configuration. On PRM800K (Appendix L) the pooled coefficient's limit error is 10.0% under the answer label and

10.0% under the human step label, the disagreement shares 0% and 20%; against the complement the grouped mean stays ahead of pooling from 32 and half problems under the pairwise choice, and under best of n best of n on the labelled problems stays ahead of the grouped mean from 96 problems under the answer label and of the pooled coefficient from half.

From eight clusters on. The budgets of the grid from eight labelled clusters to half, under best of n, on eight sets, the primary configurations, WMT23, RewardBench 2 with and without its authors' models, its one- and two-per-family panels averaged over their draws, PRM800K, and PPE in points of best-of-16 accuracy on the fixed evaluation set, give 66 point estimates of a set mean, without intervals, of which the within-cluster Pearson coefficient's is below zero by more than a twentieth of a point on 7, lowest at -0.3 on the one-per-family RewardBench 2 panels, and the grouped mean's on 32, lowest at -7.6 on PRM800K. Below eight clusters the within-cluster coefficient loses up to a point on WMT23 and RewardBench 2 and the grouped mean up to three. The PDP coefficient of DiIanni & Deutsch (2025), the Pearson correlation of the pairwise differences inside a cluster, is r_W wherever the clusters of a configuration are of one size, which they are on 26 of the 29 configurations of the primary set, WMT23, RewardBench 2 and PRM800K; run as a rule of its own on the main run's draws (`scripts/decision/pdp_rule.py`), on the other three (FRANK and the two PRM800K labels) it selects the evaluator r_W selects on at least 96% of draws at eight units and on every draw at half, and the two are never more than 0.14 points of recovery apart, so every r_W number of this paper can be read as PDP's. Tables whose captions name the held-out script's 500 splits come from that run's own draws (Appendix J); the grouped mean's human-label cell on PRM800K is 1.6 points lower there than on the main run's draws.

L Same-Model Candidates: PRM800K

The decision the within-cluster component is wanted for most directly is choosing among several samples of one model for one prompt, and no benchmark in the census releases evaluator scores on such a set (Appendix W). This appendix builds one, under a design written down before any evaluator was scored and released as `scripts/same_model/DESIGN.md`; the analysis is that document run once. *Data.* PRM800K (Lightman et al., 2024), phase-2 test split: 2762 solutions to 458 MATH problems, all sampled from one generator, GPT-4, six per problem for 420 of them, each solution's steps rated $+1$, 0 or -1 by a human labeler who stops at the first step rated -1; every problem enters with every solution the release holds for it. The problem is the cluster and its solutions the candidates. Two labels give two primary configurations: the final answer graded against the ground truth with the release's own grader, one or zero, which is the verifiable label and the reward-model use case; and the share of a solution's steps before the first step rated -1, one when none is, which is the graded human label. Solutions whose labeler gave up, 98, carry no human label and are dropped from that configuration. The two labels agree on 92% of solutions, and 0.24 of answers are correct, a share set by the release's active-learning selection of convincing wrong answers and not the generator's base rate. *Panel.* Six public reward models of four model families, fixed before scoring: the process reward models Qwen2.5-Math-PRM-7B, math-shepherd-mistral-7b-prm and Llama3.1-8B-PRM-Deepseek-Data, and the outcome reward models Skywork-Reward-V2-Llama-3.1-8B, internlm2-7b-reward and ArmoRM-Llama3-8B-v0.1. Each scores a solution from the problem statement and the solution text alone, with no label, no problem identifier and no other solution visible; every model is loaded in 8-bit so that the process never holds more than half of a shared GPU, and the same quantisation applies to all six. A process reward model's solution score is the minimum over its step probabilities, as fixed in the design: the minimum is what Math-Shepherd uses (Wang et al., 2024a), while the Qwen2.5-Math-PRM paper scores by the product and prefers the last step for models trained on Monte Carlo estimates (Zhang et al., 2025), and the release of Llama3.1-8B-PRM-Deepseek-Data scores by the mean[1], so the product, the mean and the last step are recorded and reported in the release. 1 of the six could not be loaded, ArmoRM, whose released code imports a symbol that the installed `transformers` 4.56 no longer defines; under the design it is dropped and not replaced, so the panel is five models of four families. 1 solution was dropped for every evaluator because one process reward model could not locate its step boundaries. Two environment notes: the Mistral tokenizer

[1] `https://github.com/RLHFlow/RLHF-Reward-Modeling/tree/fc39179f5a9a4dd75047a4c0311672905b9d9a04/math-rm`, its README and `prm_evaluate.py`.

Table 12: The within-cluster Pearson coefficient and the grouped Spearman mean against the pooled coefficient, by configuration: the paired difference in points of best-of-n recovery between the selected evaluators, at eight labelled clusters and at half, the estimate ± 1.96 corpus-bootstrap standard deviations of 300 splits; each set's mean, and its mean without the configuration whose value is largest in size for that rule and budget. The set means quoted in Sections 5 and 6 sit on these.

	eight clusters		half the clusters	
Configuration	$r_W - r_T$	grp $- r_T$	$r_W - r_T$	grp $- r_T$
Primary				
so762 accuracy	+1.1 [−1.3, +3.4]	+0.4 [−2.5, +3.2]	+0.0 [−0.9, +0.9]	+0.0 [−1.1, +1.1]
so762 fluency	+0.3 [−1.5, +2.1]	−0.0 [−1.9, +1.8]	+0.5 [−5.0, +6.0]	+2.9 [−7.3, +13.2]
so762 prosodic	−2.0 [−4.8, +0.9]	−2.9 [−6.8, +1.1]	−2.8 [−8.4, +2.7]	+4.7 [−2.4, +11.9]
so762	+3.1 [−0.8, +6.9]	+2.5 [−1.6, +6.6]	+0.1 [−3.0, +3.2]	+0.0 [−3.8, +3.8]
FRANK factuality	+4.1 [+0.2, +7.9]	+2.0 [−2.0, +6.0]	+6.4 [−4.2, +17.0]	+6.4 [−4.1, +17.0]
HANNA coherence	−1.0 [−4.4, +2.4]	+2.3 [−1.3, +5.9]	−3.5 [−9.8, +2.8]	+0.6 [−5.7, +6.9]
HANNA complexity	+1.8 [−2.5, +6.1]	+2.9 [−0.5, +6.4]	+3.6 [−5.8, +13.0]	+4.8 [−3.3, +13.0]
HANNA empathy	−1.8 [−4.3, +0.6]	−2.7 [−5.8, +0.4]	−0.9 [−5.5, +3.7]	−2.5 [−7.9, +2.9]
HANNA engagement	+0.9 [−2.5, +4.3]	+1.2 [−2.2, +4.5]	−4.0 [−11.3, +3.3]	−2.3 [−9.4, +4.8]
HANNA relevance	+14.9 [+4.5, +25.4]	+10.5 [+2.0, +19.0]	+22.9 [+6.8, +39.0]	+21.0 [+3.4, +38.5]
HANNA surprise	−0.5 [−3.9, +3.0]	−0.2 [−3.5, +3.1]	−1.4 [−9.2, +6.4]	+0.3 [−6.9, +7.6]
RealSumm	+3.5 [−0.4, +7.3]	+3.3 [−0.3, +7.0]	+1.7 [−4.3, +7.6]	+2.4 [−4.0, +8.7]
SummEval coherence	+0.7 [−0.8, +2.1]	−2.4 [−5.4, +0.5]	+0.0 [−2.0, +2.1]	−7.3 [−13.7, −0.8]
SummEval consistency	−0.2 [−0.9, +0.4]	+2.4 [−1.8, +6.6]	+0.0 [−0.8, +0.9]	+4.8 [−1.6, +11.1]
SummEval fluency	−0.3 [−1.5, +0.8]	−4.0 [−7.5, −0.6]	−1.3 [−3.9, +1.2]	−6.3 [−13.2, +0.5]
SummEval relevance	+1.4 [−1.6, +4.3]	−0.4 [−4.2, +3.3]	+0.0 [−0.2, +0.2]	+0.0 [−0.5, +0.5]
WMT22 ende	+0.9 [−0.8, +2.7]	+2.0 [−0.3, +4.3]	+0.2 [−2.1, +2.6]	−0.0 [−4.2, +4.2]
WMT22 enru	+4.1 [+1.8, +6.3]	+5.3 [+2.9, +7.7]	+1.5 [−1.6, +4.6]	+1.0 [−2.9, +5.0]
WMT22 zhen	+0.6 [−0.6, +1.9]	−0.1 [−1.6, +1.4]	+0.3 [−1.4, +1.9]	+0.2 [−1.9, +2.2]
mean	+1.6 [+0.7, +2.6]	+1.2 [+0.1, +2.2]	+1.2 [−0.5, +2.9]	+1.6 [−0.2, +3.5]
mean without the largest configuration	+0.9 [+0.1, +1.8]	+0.6 [−0.3, +1.6]	+0.0 [−1.6, +1.6]	+0.5 [−1.2, +2.3]
WMT23 MQM				
WMT23 ende	+0.5 [−1.4, +2.3]	−0.5 [−3.4, +2.4]	−0.7 [−4.7, +3.3]	−3.2 [−8.9, +2.4]
WMT23 heen	+0.2 [−1.4, +1.9]	−0.0 [−2.1, +2.1]	+0.5 [−3.0, +4.0]	−3.3 [−8.7, +2.2]
WMT23 zhen	−1.0 [−2.7, +0.7]	+0.4 [−1.7, +2.4]	+2.4 [−0.4, +5.2]	+4.7 [−0.0, +9.4]
mean	−0.1 [−1.1, +0.9]	−0.0 [−1.4, +1.3]	+0.7 [−1.4, +2.8]	−0.6 [−3.6, +2.4]
mean without the largest configuration	+0.4 [−0.8, +1.5]	+0.2 [−1.3, +1.7]	−0.1 [−2.9, +2.6]	−3.2 [−7.3, +0.8]
RewardBench 2				
RB2 factuality	+0.7 [−1.4, +2.8]	−0.7 [−2.9, +1.5]	+0.9 [−2.6, +4.4]	+0.8 [−2.9, +4.5]
RB2 precise if	+4.3 [−0.3, +8.8]	+4.0 [−1.1, +9.1]	+0.0 [−2.4, +2.4]	+0.0 [−2.7, +2.7]
RB2 math	+0.5 [−1.8, +2.8]	−2.0 [−5.1, +1.0]	+4.0 [−0.5, +8.5]	+7.5 [+0.7, +14.3]
RB2 safety	−0.2 [−1.1, +0.8]	−4.9 [−6.7, −3.2]	+1.5 [−1.0, +4.1]	+1.5 [−1.4, +4.5]
RB2 focus	−0.6 [−1.4, +0.2]	−8.2 [−10.3, −6.1]	+0.0 [−0.0, +0.0]	−0.6 [−1.7, +0.4]
mean	+0.9 [−0.2, +2.1]	−2.4 [−3.8, −1.0]	+1.3 [−0.1, +2.6]	+1.8 [+0.2, +3.5]
mean without the largest configuration	+0.1 [−0.7, +0.9]	−0.9 [−2.6, +0.8]	+0.6 [−0.6, +1.8]	+0.4 [−1.0, +1.8]
PRM800K				
PRM800K answer correct	+0.3 [−1.6, +2.3]	−8.4 [−11.5, −5.3]	+0.0 [+0.0, +0.0]	+0.0 [−0.5, +0.5]
PRM800K human progress	−0.3 [−2.5, +1.9]	−6.2 [−9.2, −3.2]	+0.0 [−0.2, +0.2]	−0.4 [−5.7, +4.9]
mean	+0.0 [−1.4, +1.5]	−7.3 [−9.6, −5.0]	+0.0 [−0.1, +0.1]	−0.2 [−2.9, +2.5]
mean without the largest configuration	−0.3 [−2.5, +1.9]	−6.2 [−9.2, −3.2]	+0.0 [−0.2, +0.2]	+0.0 [−0.5, +0.5]

Table 13: The same-model checks: held-out selection error, percent of evaluator pairs, of four rules computed on half the problems or prompts and scored on the other half, 500 splits, and the percentage of splits on which the one evaluator the rule selects is the best on the evaluation half (on PPE, by the accuracy of the top-scored response; both: the two payoffs averaged, each in Table 17). Ev. is the panel size; bold the lowest error in the row. Appendices L and M.

			Selection error, %				Selects best, %		
Configuration	Payoff	Ev.	r_T	r_W	grp	dec	r_T	grp	dec
PRM800K, GPT-4 solutions of one MATH problem, five reward models fixed before scoring									
answer correct	pairwise	5	12.3	9.5	**5.6**	6.6	100	100	100
	best of n	5	7.7	**5.9**	10.3	7.5	100	100	100
human step label	pairwise	5	9.2	8.5	**7.2**	7.6	99	97	96
	best of n	5	13.6	13.6	12.1	**9.5**	100	98	100
PPE, responses of one model to one prompt, the panel of Frick et al. (2025) as far as it loads, with their length baseline									
MMLU-Pro	both	6	19.6	**19.0**	22.9	22.5	68	24	41
MATH	both	6	17.2	**16.0**	18.3	22.9	22	16	23
GPQA	both	6	**28.8**	30.4	30.1	32.8	33	20	18
IFEval	both	6	21.8	20.0	**15.6**	19.7	89	66	76
MBPP-Plus	both	6	22.1	17.9	15.4	**15.1**	98	100	100

needed `protobuf`, and Qwen2.5-Math-PRM-7B's forward is called with its cache disabled because its released code uses a cache interface that version has changed; neither touches a weight or a score. *Analysis.* The machinery of the held-out experiment, unchanged: 500 random half-splits of the problems, every pair of evaluators, the pooled and within-cluster coefficients, the grouped Spearman mean and the decision itself computed on the selection half, the pairwise choice and best of n scored on the evaluation half, top-1 regret, and the cross-utility matrix of the same statistics. The primary comparison, fixed in advance, is the grouped mean against the pooled coefficient, cell-equal over both labels and both decisions.

Result. Between problems lie 29% of the variation of answer correctness, an intraclass correlation of 0.15, and 41% of the step-progress label; the leave-one-out label oracle of Proposition 3 reaches a pooled correlation of 0.26 and 0.44 without reading a solution. The 5 evaluators' pooled correlations with answer correctness run from 0.28 to 0.62 and their within-problem correlations from 0.17 to 0.58, and they recover between 15% and 64% of the available best-of-n gain. Choosing between them on half the problems and paying on the other half, the pooled coefficient errs on 10.0% of evaluator pairs under the answer label and 11.4% under the human label, the within-cluster coefficient on 7.7% and 11.1%, the grouped mean on 7.9% and 9.7%, and the decision itself on the selection half on 7.1% and 8.5%; by decision, the grouped mean errs on 5.6% and 10.3% against 12.3% and 7.7% for pooling under the answer label, and the grouped mean is ahead of pooling on 3 of the four cells. Named one evaluator under best of n, every rule selects the same model, the strongest of the five, and gives up none of the span between the best evaluator and the average one; the errors above are on the other pairs. Under the answer label alone and the two payoffs, the pairwise accuracy on the selection half errs on 8.6% and best of n on the selection half on 8.5%, against 8.3% for the grouped mean and 10.1% for pooling. Those four are computed by the cross-utility script on 500 splits of its own, so the last two are a second estimate of the answer-label means of the primary run, 7.9% and 10.0%, and differ from them by at most 0.3 points; the two scripts draw their own splits and the held-out one keeps its own conventions (Appendix J), and on the human-label pairwise cell they differ by more, the grouped mean at 8.8% on the main run's draws (Table 11) against 7.2% in Table 13. What this cannot show is stated with it: one generator, one task family, one split of one corpus and a panel of six we chose, a same-model check on a verifiable task and not a survey, and its candidates were selected by the release's procedure rather than sampled from the generator's distribution.

Definitions. The human step label of a solution is the number of its steps before the first step rated -1, divided by the number of steps of the released solution, rated or not, since the labeler stops at the first -1 and every later step counts as after the error; it is zero when the first step is rated -1, one when no step is, and absent for the 98 solutions whose labeler gave up. The 92% agreement is exact agreement between the binary no-error label and the answer label over the labelled solutions. *Sensitivity, and what gets named.* `scripts/same_model/sensitivity_prm800k.py` reruns the analysis three ways, none of which replaces the design. With the process reward models aggregated by the last step instead of the minimum, the pooled coefficient errs on 39.1% of evaluator pairs and the grouped mean on 15.6%, by the product 12.6% against 8.7%, by the mean 14.0% against 12.0%, against

Table 14: The PRM800K panel in sample, on every problem: each reward model and, for the process reward models, each aggregation of its step probabilities, under the answer label and the human step label. r_T pooled Pearson, r_W within-problem Pearson, grp the mean of the per-problem Spearman correlations, pair the pairwise accuracy inside problems and bon the best-of-n recovery, both in percent. The minimum is the aggregation the design fixed; the other three are the sensitivity of Appendix L.

			Answer correct					Human step label				
Model	Kind	Agg.	r_T	r_W	grp	pair	bon	r_T	r_W	grp	pair	bon
Qwen2.5-Math-PRM-7B	PRM	min	0.62	0.58	0.49	82	64	0.55	0.47	0.32	63	52
math-shepherd-mistral-7b-prm	PRM	min	0.33	0.23	0.14	60	19	0.33	0.27	0.19	58	22
Llama3.1-8B-PRM-Deepseek-Data	PRM	min	0.44	0.36	0.35	72	39	0.41	0.27	0.22	59	33
Skywork-Reward-V2-Llama-3.1-8B	ORM	–	0.28	0.17	0.17	61	15	0.34	0.20	0.16	56	16
internlm2-7b-reward	ORM	–	0.44	0.34	0.30	69	34	0.46	0.35	0.26	61	30
Qwen2.5-Math-PRM-7B	PRM	last	0.32	0.27	0.47	80	56	0.19	0.13	0.22	59	37
Qwen2.5-Math-PRM-7B	PRM	product	0.65	0.62	0.49	81	64	0.60	0.51	0.36	66	53
Qwen2.5-Math-PRM-7B	PRM	mean	0.47	0.43	0.52	83	67	0.55	0.47	0.41	68	55
math-shepherd-mistral-7b-prm	PRM	last	0.33	0.25	0.20	63	24	0.33	0.28	0.22	59	25
math-shepherd-mistral-7b-prm	PRM	product	0.33	0.24	0.10	57	17	0.31	0.23	0.20	59	22
math-shepherd-mistral-7b-prm	PRM	mean	0.33	0.19	0.17	62	23	0.38	0.32	0.24	61	27
Llama3.1-8B-PRM-Deepseek-Data	PRM	last	0.28	0.19	0.33	71	30	0.19	0.10	0.16	57	19
Llama3.1-8B-PRM-Deepseek-Data	PRM	product	0.50	0.39	0.31	70	35	0.48	0.35	0.24	60	31
Llama3.1-8B-PRM-Deepseek-Data	PRM	mean	0.36	0.24	0.40	75	44	0.42	0.27	0.27	61	38

Table 15: PPE against a fixed evaluation set: five rules computed on b distinct prompts of the 96-prompt pool, 500 draws, scored on 192 further prompts a set no rule uses, five sets with equal weight. Error: percent of evaluator pairs on which the rule selects the evaluator that is worse there. Accuracy: the release's best-of-16 accuracy there of the evaluator the rule selects, beside the best and the mean evaluator. Bold: best in the block; paired differences with intervals: Table 16.

	Error, pairwise choice					Error, best of n					Accuracy of the selected evaluator						
b	r_T	r_W	grp	pair	bon	r_T	r_W	grp	pair	bon	r_T	r_W	grp	pair	bon	best	mean
8	29.8	**26.8**	28.1	28.1	37.0	32.4	**28.1**	29.3	29.1	37.5	55.1	**56.7**	56.6	56.6	55.0	58.9	50.5
24	21.9	**19.8**	20.6	20.7	31.6	26.0	**21.9**	22.6	22.5	32.2	56.6	**57.9**	57.3	57.3	56.1	58.9	50.5
48	17.5	**16.2**	16.8	16.8	27.8	22.4	**18.6**	19.1	19.0	28.2	57.4	**58.3**	57.5	57.5	56.5	58.9	50.5
96	**12.0**	13.3	14.7	13.3	24.7	17.3	16.0	17.3	**16.0**	24.7	57.6	**58.8**	57.5	57.6	56.9	58.9	50.5

10.7% and 8.8% under the minimum, which reproduces the primary run; the grouped mean is ahead on 4, 3 and 3 of the four cells. Named one evaluator, every rule selects Qwen2.5-Math-PRM-7B on every split under the answer label, and all four rules select the same model on 97% and 98% of splits under the human label, where it is the best evaluator on the evaluation half on 99% and 100% of splits; on this panel the evaluator choice is settled by every rule and the selection errors above are on the other pairs. Treating every aggregation of every process reward model as an evaluator of its own, 14 evaluators, the pooled coefficient errs on 24.6%, the within-cluster one on 22.8%, the grouped mean on 11.8% and the decision on the selection half on 10.8%, the grouped mean ahead of pooling on 4 of the four cells and giving up 2% of the span against 16% for pooling; by label, 27.9% against 7.3% under the answer label and 21.3% against 16.4% under the human one. Table 14 gives the in-sample statistics of every model and aggregate under both labels.

M Same-Model Candidates: PPE

The second same-model check uses a candidate set and an evaluator panel that were not ours to choose, under a design written before any response was scored and released as `scripts/same_model/DESIGN_PPE.md`. *Data.* The five verifiable-correctness sets of Preference Proxy Evaluations (Frick et al., 2025), MMLU-Pro (Wang et al., 2024b), MATH (Hendrycks et al., 2021), GPQA (Rein et al., 2024), IFEval (Zhou et al., 2023) and MBPP-Plus of the EvalPlus project (Liu et al., 2023a): 512 prompts a set, 507 for MBPP-Plus, 32 responses to each prompt all sampled from one of four generators, GPT-4o-mini, Llama-3-8B-Instruct, Claude-3-Haiku and Gemma-2-9B-it, and a correctness label per response from the release's checker. From each set 96 prompts are drawn uniformly at random with a fixed seed from the rows sorted by their identifier, and the first 16 responses of each are taken by column index, 7680 responses in all; the prompt is the

cluster, its responses the candidates, the label is correctness, and a prompt whose labels are all equal asks no question. Responses longer than 12,000 characters are cut at the end for every evaluator alike, one response in the sample. *Panel.* The rule, fixed before scoring: every reward model of PPE's own panel whose weights are public, which loads through a `transformers` Auto class with or without the repository's own code, and which fits in half of a 24 GB GPU in 8-bit, attempted in a fixed order, internlm2-7b-reward, internlm2-1_8b-reward, Skywork-Reward-Llama-3.1-8B, Llama-3-OffsetBias-RM-8B, Eurus-RM-7b and ArmoRM-Llama3-8B-v0.1, plus PPE's length baseline, the number of characters of the response; Athene-RM-8B and Starling-RM-7B-alpha are excluded by the loading clause, the 20B to 340B models by the memory clause, and the API judges by not being reward models we can run. A model that fails to load or to score is dropped and not replaced. Five of the six loaded and scored every response but 16, which one model found too long and which are dropped for every evaluator, leaving 7664. The sixth, ArmoRM, produced scores in this run, and they are left out: it loaded only because the copy of its released code in our model cache carried a one-line edit, a docstring constant the installed `transformers` no longer defines, made during the PRM800K round and not adopted (Appendix L), and the rule admits the repository's own code, not a local edit of it; its scores ship in the release, enter no panel, and the cached file was restored to the repository's version, so the panel is 6 evaluators with the length baseline. Each model sees the released prompt as the user turn and the response as the assistant turn in its own chat template, Eurus in its bracketed format, with no label, no prompt identifier and no other response; models above 3B parameters are loaded in 8-bit and the 1.8B model in bf16, and weights downloaded for the experiment were removed after scoring. *Analysis.* The machinery of Section 4, unchanged, on five configurations clustered by prompt, 500 half-splits of the prompts each, both decisions, and the label-budget curves of Appendix K; and, in the release's own units, the best-of-16 accuracy on the evaluation half, the share of prompts whose top-scored response is correct, of the evaluator each rule selects on the selection half, against the best and the mean evaluator of the panel there. The primary comparison, fixed in advance, is the grouped mean against the pooled coefficient, the five sets with equal weight. *Result.* Between prompts lie 24% to 38% of the label variation by set. Over the five sets and both decisions the pooled coefficient errs on 21.9% of evaluator pairs, the within-cluster one on 20.7%, the grouped mean on 20.4% and the decision on the selection half on 22.6%, the grouped mean ahead of pooling on 4 of the ten cells and 2 of the five sets; by decision, 19.2% against 16.9% under the pairwise choice and 24.6% against 23.9% under best of n; without the length baseline, 27.0% against 24.6%, the grouped mean ahead on 5 cells. Named one evaluator, the one the grouped mean names is the best of the panel by best-of-16 accuracy on the evaluation half on 45% of splits, against 62% for the pooled coefficient, 49% for the within-cluster one and 58% for best of n on the selection half; those evaluators reach a best-of-16 accuracy of 61.7%, 63.5%, 62.1% and 62.5% there, against 66.2% for the best evaluator, 52.6% for the mean one and 37.0% for the worst; under the paper's own decisions the pooled coefficient's regret is 22% and 27% of the span against 31% and 42% for the grouped mean. So the primary prediction holds over evaluator pairs and not for the one evaluator named, and both are reported. Along the budget against the measurement set (Table 11) the within-cluster coefficient is the best of the five rules at 9 of the 9 common budgets under the pairwise choice and the pooled coefficient at 0; the pooled coefficient's limit error on the pool is 9.3%, and 9% of evaluator pairs are ordered one way by it and the other by the within-prompt coefficient, none on MATH and GPQA and most on IFEval, where grouping pays most. Table 13 carries every set and Table 17 every cell. *A fresh pool.* The intervals of Tables 15 and 16 resample the measurement set's prompts with the pool and the selection draws held fixed. `scripts/same_model/ppe_pool_boot.py` resamples both, 200 times: the pool's prompts with replacement, a prompt weighted by its multiplicity as in the corpus bootstrap, 300 selection draws on the resampled pool (a random half of the present prompts, then b distinct prompts of it), every rule computed on the drawn prompts, and the selected evaluator scored on an independent resample of the measurement set; the five tasks are resampled independently. A replicate of the pool reads b distinct present prompts of a random half of the resampled pool with their multiplicities as weights, the convention of every corpus bootstrap here (Section 4). Over the five tasks the paired accuracy difference against pooling at eight prompts is +1.6 for the within-cluster coefficient, $[+1.1, +2.1]$ with the measurement set alone resampled and $[+0.5, +2.3]$ with both, 1.9 times as wide, both percentile intervals, and +1.5 for the grouped mean, $[+0.9, +2.0]$ and $[+0.4, +2.5]$; at 24 prompts +1.3 $[+0.0, +3.0]$ and +0.7 $[-0.6, +2.6]$, and at 48 +0.9 $[-0.1, +3.0]$ and +0.1 $[-0.8, +2.6]$; best of n computed on the pool is -0.1 $[-1.6, +0.7]$ at eight. So the eight-prompt gain of the coefficient rules survives a fresh pool and the 24- and 48-prompt gains do not clear zero under one. *What this cannot show.* 96 prompts a set, one binary label per response and no graded human label, candidates of one prompt from one

generator but four generators across prompts, prompts chosen by the release, which kept those whose 32 responses were neither all right nor all wrong, and a panel of at most six reward models and a length baseline: a same-model check on verifiable tasks with a panel fixed by others, not a survey.

A fixed measurement set. The addendum to the design, written on 15 September 2026 before any further response was scored, holds 192 further prompts a set as a measurement set that no rule uses: drawn uniformly without replacement from the rows not in the 96-prompt pool, seed fixed, the first 16 responses of each, the same cut, scored by the five panel models, 15360 responses of which 70 were too long for one model and are dropped for every evaluator; 4 to 28 prompts a set carry sixteen equal labels and enter no within-prompt statistic, leaving 164 to 188 that do. `scripts/decision/targets.py` reads the five rules off b distinct prompts of the selection half of the pool, 500 half-splits, and the whole pool at 96, and pays every evaluator pair by the decision on the measurement set, by the decision on the pool's other half and by the decision on the pool itself, on the same draws; the one evaluator each rule selects is paid its decision score and the release's best-of-16 accuracy on the measurement set. Table 15 gives the budgets of 8, 24, 48 and 96 prompts against the measurement set and Table 16 every difference against the pooled coefficient with a 2.5 to 97.5 percentile interval over 1000 cluster bootstraps of the measurement set's prompts, the draws held fixed. Over the 10 budgets the grouped mean is ahead of pooling on 9 under the pairwise choice and 10 under best of n, the within-cluster coefficient on 10 and 10, and best of n on the pool is the worst of the five on 10 and 10; the pooled coefficient is the best rule at no budget up to half the pool under either payoff, and at the whole pool, 96 prompts, it is the best under the pairwise choice (Table 15). Two independent draws of 48 prompts make the rules select different evaluators of a pair on 12.3%, 11.2%, 12.1%, 12.7% and 25.2% of pairs. The measurement set's own uncertainty is reported two ways: the share of decisive pairs whose ordering holds in at least 95% of 1000 bootstrap resamples of its prompts, 33% to 93% by set under the pairwise choice and at least 33% under best of n (`analyse_ppe_fixed.py`); and the paired intervals of Table 16, which at 48 prompts include zero for every difference against pooling among the four coefficient rules and exclude it for best of n on the pool under the pairwise choice, +10.3 points $[+5.4, +12.9]$, and at eight prompts exclude it for the within-cluster coefficient's error, -3.0 $[-4.3, -1.1]$, and for the accuracy of the evaluators the within-cluster coefficient, the grouped mean and the pairwise accuracy name, +1.6, +1.5 and +1.5 points above pooling's. The pool's other half, 48 prompts, is a noisier evaluation side than the measurement set and reads the coefficient rules higher and best of n on the pool lower, +1.5, +0.3 and -2.9 points under the pairwise choice, the compression of Appendix J from both sides. By set at 48 prompts, under the pairwise choice, the grouped mean errs on 18.4%, 6.6%, 35.4%, 20.6% and 3.1% against 10.3%, 5.8%, 32.9%, 24.2% and 14.4% for pooling on MMLU-Pro, MATH, GPQA, IFEval and MBPP-Plus, and the evaluator it selects reaches a best-of-16 accuracy of 59.8%, 49.7%, 43.4%, 59.5% and 75.0% against 61.6%, 49.4%, 43.4%, 57.8% and 74.8% for the pooled coefficient's, with the best evaluator at 61.7%, 50.3%, 43.9%, 63.8% and 75.0%. By generator of the measurement prompts, 18 to 87 prompts a cell over 20 generator-by-set cells, the evaluator the pooled coefficient names at 48 prompts is ahead of the grouped mean's in best-of-16 accuracy on 6 cells and behind on 11, by at most 5.9 points (`analyse_ppe_fixed.py`, whose draws are from the whole pool). The results on the pool's other half alone, Table 17 and the PPE rows of Table 13, are the held-out experiment of Appendix H on this panel.

N The Candidate-Count Manipulation

Freeze, and what it certifies. The protocol of Section 6 (`prereg/PROTOCOL_candidates.md`) was frozen together with the analysis script (`scripts/decision/candidate_count.py`), the grading script (`candidate_report.py`) and the fifteen project modules they import or replay: their SHA-256 are in `prereg/FREEZE_candidates.txt`, which this machine wrote at 2026-09-18 22:22:21 UTC, and an RFC 3161 token over that record, issued by a public time-stamping authority whose clock read 22:21:58, is kept beside it with the authority's certificates and the command that verifies it. The token proves that the record, and so the predictions and the code, existed by that time; that no selection rule or utility had been computed on a subsampled cluster of real data before it is the authors' statement, which the file times of the release and one contradicted prediction are consistent with and cannot prove. Three things the protocol's list of what was known leaves out or understates. First, the dry run: before the freeze the script ran on two synthetic worlds generated inside it, where its every-candidate path was checked against `targets.run_cell`,

Table 16: PPE, paired differences against the pooled coefficient on the fixed evaluation set, in points, five sets with equal weight. The point estimate is the difference on the evaluation set itself, the same quantity as Table 15; the interval is the 2.5 to 97.5 percentile of the five-set mean over 1000 cluster bootstraps of each set's measurement prompts, drawn independently for the five sets, conditional on the pool and on the selection draws, so it is the uncertainty of the evaluation set and not of the pool. Error over evaluator pairs under each payoff (negative: the rule errs less than pooling); best-of-16 accuracy of the selected evaluator (positive: the rule's evaluator earns more).

b	Rule	Error, pairwise choice	Error, best of n	Accuracy of the selected evaluator
8	r_W	−3.0 [−4.3, −1.1]	−4.4 [−5.6, −1.6]	+1.6 [+1.1, +2.1]
	grp	−1.7 [−3.3, −0.1]	−3.1 [−4.5, −0.7]	+1.5 [+0.9, +2.0]
	pair	−1.7 [−3.3, −0.1]	−3.3 [−4.7, −0.9]	+1.5 [+1.0, +2.0]
	bon	+7.2 [+5.5, +8.5]	+5.0 [+1.5, +6.8]	−0.1 [−0.6, +0.4]
24	r_W	−2.1 [−4.9, +1.2]	−4.1 [−6.4, +0.1]	+1.3 [+0.6, +2.0]
	grp	−1.3 [−4.5, +1.4]	−3.4 [−5.8, +0.3]	+0.7 [+0.0, +1.4]
	pair	−1.1 [−4.4, +1.6]	−3.5 [−6.1, +0.2]	+0.7 [+0.1, +1.4]
	bon	+9.7 [+6.3, +11.7]	+6.2 [+0.7, +8.8]	−0.5 [−1.2, +0.2]
48	r_W	−1.3 [−5.3, +2.9]	−3.9 [−6.7, +1.3]	+0.9 [−0.1, +1.9]
	grp	−0.7 [−5.9, +2.6]	−3.3 [−6.7, +1.7]	+0.1 [−0.8, +1.1]
	pair	−0.7 [−5.7, +2.8]	−3.5 [−6.6, +1.6]	+0.1 [−0.7, +1.0]
	bon	+10.3 [+5.4, +12.9]	+5.8 [−1.2, +9.0]	−0.9 [−1.9, +0.2]
96	r_W	+1.3 [−4.0, +6.7]	−1.3 [−6.7, +4.2]	+1.2 [−0.2, +2.7]
	grp	+2.7 [−5.3, +8.0]	−0.0 [−6.9, +5.3]	−0.1 [−1.5, +1.5]
	pair	+1.3 [−6.7, +6.7]	−1.3 [−6.6, +4.1]	+0.0 [+0.0, +0.0]
	bon	+12.7 [+4.7, +15.3]	+7.3 [−4.8, +12.7]	−0.7 [−2.3, +0.9]

Table 17: The PPE sets of Table 13 under each payoff: held-out selection error of four rules and the percentage of splits on which the evaluator each rule selects is the best of the panel by best-of-16 accuracy on the evaluation half; dec is the payoff's own statistic on the selection half.

			Selection error, %				Selects best, %		
Configuration	Payoff	Ev.	r_T	r_W	grp	dec	r_T	grp	dec
MMLU-Pro	pairwise	6	22.3	**19.8**	24.7	24.9	68	24	28
	best of n	6	**16.9**	18.2	21.2	20.2	68	24	53
MATH	pairwise	6	11.4	**9.9**	12.7	13.1	22	16	16
	best of n	6	22.9	**22.1**	23.9	32.7	22	16	30
GPQA	pairwise	6	**26.0**	29.1	27.3	27.1	33	20	24
	best of n	6	**31.7**	31.8	32.8	38.5	33	20	12
IFEval	pairwise	6	21.7	19.4	**14.3**	15.7	89	66	58
	best of n	6	21.9	20.7	**16.9**	23.7	89	66	94
MBPP-Plus	pairwise	6	14.4	8.1	**5.8**	5.9	98	100	100
	best of n	6	29.7	27.7	25.0	**24.3**	98	100	100

`boot300.corpus_bootstrap_batched` and `ppe_pool_boot.run_task`, and a small full run was graded five and a half minutes before the freeze; its six contrasts came out in the directions the protocol then predicted, C1 $+0.9$ $[-1.2, +2.9]$, C2 -0.5 $[-2.3, +1.3]$, C3 $+1.5$ $[+0.1, +2.9]$, C4 $+1.0$ $[-0.1, +2.0]$ in the world built like PPE and C5 $+4.9$ $[+0.3, +9.5]$, C6 $+3.7$ $[+0.7, +6.7]$ in the other, where the protocol says only that the grouped mean's gain shrank with the candidates and the within-cluster coefficient's shrank less; both files are released (`prereg/dry_run/`). Second, the power analysis of Appendix Z, of 13 September, subsamples the utterances of each speechocean762 speaker to scale standard errors of correlation differences; it computes no selection rule and no utility, and the protocol's "no statistic on a subsampled cluster, by this script or any other" should have read "no selection rule or utility". Third, the protocol's statement of what its authors had seen, every number of the draft then current, leaves out what was computed between that draft and the freeze, all of it with every candidate kept: the 300-split intervals, the three recalibrated worlds, the leave-one-out intervals and the family-panel summary. On the real data the script had been run with every candidate kept only, which reproduced the released files: point estimates to 10^{-16}, corpus-bootstrap replicates to the seven decimals the files store, both PPE intervals exactly. The protocol writes of Appendix S for the coverage of the interval; that appendix is Appendix AA here.

Design. Every labelled cluster keeps $m \in \{2, 3, 4, 6, 8\}$ of its candidates, drawn uniformly without replacement, a cluster with at most m keeping all (FRANK's clusters hold five at the median, HANNA's ten, the others twelve to twenty-four, PPE's sixteen). There are 20 independent subsamples per configuration and m; split s of the main run's draws, and split s of every bootstrap replicate,

uses subsample s modulo 20, and every rule on a draw sees the same subsample, so the intervals are conditional on those 20 subsamples. The rules are the pooled and within-cluster coefficients, the grouped mean, the two decision statistics computed on the labelled clusters and the equal-weight mean of per-cluster Pearson coefficients, the grouped Pearson cell of Table 6; the benchmark's own statistic is not run. The evaluation side is untouched. On PPE a draw's m responses are a random sample of one model's, which is what a practitioner who samples m would hold; on the exploratory configurations and WMT23 a cluster's candidates are the outputs of different systems and a practitioner who labels fewer of them would label the same systems in every cluster, which the random subsets of the manipulation are not (a variant that does is below). On those sets the interval is the corpus bootstrap of Section 4, its streams consumed in the same order so that replicate k of every m is the same resampled corpus and the same splits; on PPE it is the bootstrap of the pool and the evaluation set together of Appendix M, 200 replicates of 300 selection draws where the estimate averages 500, the estimate ± 1.96 standard deviations, which the protocol fixed as the grading interval, with the percentile interval and the percentile interval of the evaluation set alone beside it. When the protocol was graded none of these intervals had been run through a calibration; those of the exploratory configurations have been since (below), and those of PPE, of the addendum's panels and of WMT23 have not; the bootstrap mean of a level does not sit on its estimate, a replicate's eight distinct clusters weighted by their multiplicities holding fewer effective labels than the corpus's own eight: at four candidates on the exploratory configurations it is 1.4 bootstrap standard deviations below for the grouped mean and the within-cluster coefficient and 2.0 for the per-cluster Pearson mean and the pairwise accuracy, and over the 120 levels with fewer than all candidates it is beyond one standard deviation on 29, above the estimate on 9 of those, and at most 2.2 (`candidate_boot_displacement.py`). The grading interval, the estimate ± 1.96 standard deviations, is centred on the estimate, so the percentile interval is printed beside every level whose interval ends near zero; neither form has been calibrated for these levels, and no preference between them is claimed. The Monte Carlo standard error of the split average is 0.15 points for the primary mean of the grouped mean against pooling at eight clusters and 0.09 at half. Reporting deviations: the protocol lists every m and every budget of the grid among what is reported, the tables print eight clusters, 24, 48 and half, and the other budgets are in the released result files; the grading script was last saved eleven seconds after it wrote the synthetic dry-run report, an edit that was not recorded, and the hash record covers the later file, with which every real-data grade was produced (`prereg/dry_run/README.md`).

Grades. Table 20: six graded readings, PPE the primary test and the exploratory configurations the replication, no primary hypothesis among them and no correction, the grades descriptive. 3 are supported, 2 consistent and 1 contradicted. One grade differs between the intervals, as the protocol said would be reported: C1 is supported under the percentile interval of the evaluation set alone, $[+0.3, +0.9]$, which is conditional on the pool, and consistent under the interval the protocol fixed and under its percentile form. C2 is contradicted under all three; its interval clears zero by about one Monte Carlo standard error of the split average, and under one of four other seeds it would be graded not supported (Table 26), the withdrawal below following either way. The protocol had said the account would be withdrawn from the paper unless C1 to C3 were at least consistent, and it is withdrawn: the paper no longer offers the number of candidates as what turns the sign at eight units. What the manipulation supports is a fall of the grouped mean's gain at a fixed number of clusters, eight, on the exploratory configurations, and the prediction that carried it, C5, could have failed less easily than C2; C3 and C6 are differences of levels at four candidates, part of which exists with every candidate kept (+0.5 of C6's +1.5 on the exploratory configurations), and the difference in the falls, which carried no prediction, is +1.0 $[+0.3, +1.7]$ on the exploratory configurations and $+0.0$ $[-0.5, +0.5]$ on PPE. At half the clusters the fall is seen on the exploratory configurations only: on WMT23 no correlation rule's gain falls from all candidates to four, and on RewardBench 2 the within-cluster coefficient's gain rises with two completions (Tables 24 and 30). One task carries PPE's readings (Table 21): without MBPP-Plus the grouped mean at four responses is -0.3 $[-1.0, +0.4]$, C1 would be graded supported, C2 consistent and C4 consistent. Table 22 puts the sign counts and the leave-one-out beside every set mean on the released panels. Of the two levels the protocol left without a prediction, the grouped mean at four candidates is -0.97 $[-1.90, -0.04]$ against pooling on the exploratory configurations (-1.4 to -1.3 under four other seeds, the frozen seed being the mildest of five), crossing zero on 9 of the 19 leave-one-outs, and the within-cluster coefficient +0.5 $[-0.3, +1.3]$ (Table 22).

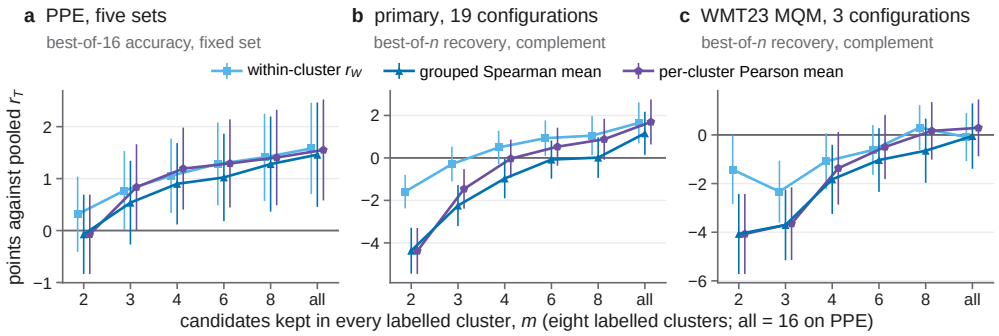


Figure 7: **The candidate-count manipulation.** Every labelled cluster keeps m of its candidates at random. Frozen in advance: six graded readings, no correction, on the grouped mean and r_W at $m = 4$ and all, on PPE and the exploratory configurations; WMT23, the per-cluster Pearson mean and the other m carried no prediction. Coverage of the exploratory configurations' intervals: Table 18; PPE's is not measured. Paired difference against the pooled coefficient computed on the same subsampled clusters, eight labelled clusters, the selected evaluator scored as in Tables 5 and 15; estimate ± 1.96 bootstrap standard deviations (Tables 19 and 24).

Every rule, budget and payoff. Table 19 prints the eight-cluster differences of Figure 7 and Table 24 every rule at both budgets under both payoffs. With two candidates a cluster's Pearson and Spearman coefficients are both ± 1 and the grouped mean, the per-cluster Pearson mean and the pairwise accuracy are one statistic, and 3.3 evaluators are tied at its top on the primary configurations against 1.0 with every candidate (Table 23); there it loses 4.4 points of recovery to pooling at eight clusters and 1.9 at half, where the within-cluster Pearson coefficient, which keeps the size of each difference, loses 1.6 and is level (+0.0). Two candidates drawn at random with their absolute labels are not a preference pair, which is chosen, carries a relative label and leaves the labels no between-cluster part, so nothing here speaks to selecting a reward model on preference data. The clusters that still ask a question of a within-cluster rule fall with m, from 7.7 of eight prompts on PPE to 5.2 at four responses and 2.9 at two. On PPE the fall is a small-budget effect: at 24 and 48 prompts the grouped mean's gain moves by less than 0.3 points between sixteen responses and four (Table 24).

Clusters falling silent, or a noisier statistic. Removing candidates does two things to a per-cluster statistic: some clusters stop asking a question, their kept candidates carrying one label value or an evaluator's kept scores one value, and in the clusters that still ask the statistic is computed over fewer values. Read after the fact, the exploratory configurations were split by it. On 9 of them (the six HANNA labels, RealSumm, SummEval coherence and relevance) four candidates leave 7.9 of eight labelled clusters asking where every candidate leaves 8.0; on the other 10 the asking clusters fall from 7.4 to 6.0. The grouped mean's gain over pooling falls by +2.1 $[+0.8, +3.3]$ on the first group, on 7 configurations of 9 and by +1.7 $[+0.5, +2.9]$ without HANNA relevance, and by +2.2 $[+1.1, +3.3]$ on the second, on 9 of 10; the difference between the groups is -0.1 $[-1.7, +1.4]$. The fall is as large where no cluster falls silent, and the split cannot say more than that: the groups were formed after the results were seen, by a threshold of 97.5% of the asking clusters kept that separates cases as close as FRANK and RealSumm, they differ in benchmark and in how many labels going from all candidates to four removes, the interval of their difference allows silent clusters to account for most of the fall, and equal falls in two groups do not separate two mechanisms that may each act in one. SummEval is the one benchmark with configurations on both sides, and there the fall is larger on the two labels that lose asking clusters (consistency and fluency) than on the two that do not. No mechanism for the fall is claimed.

The same systems in every cluster. On the exploratory configurations the frozen manipulation keeps a different random subset of candidates in every cluster, where a practitioner who labels fewer candidates would label the outputs of the same systems everywhere. Table 25, read after the fact (`candidate_systems.py`, the frozen script's draws, rules, scoring and bootstrap streams with only the subsampling replaced): 20 draws of m systems a configuration, every labelled cluster keeping the outputs of those systems, on the 15 exploratory configurations whose candidates carry a

system label. From every system to four the grouped mean's gain over pooling at eight clusters falls by +2.4 $[+1.4, +3.4]$, on 13 configurations of 15, where the random subsets give +2.1 $[+1.1, +3.1]$ on the same configurations; the within-cluster coefficient's falls by +1.5 $[+0.8, +2.3]$ against +1.0, it stays +1.3 $[+0.6, +2.0]$ above the grouped mean at four, and the grouped mean's fall exceeds it by +0.9 $[+0.0, +1.7]$. With two systems the same-systems variant costs more than two random candidates at half the clusters, -4.6 $[-6.4, -2.8]$ for the grouped mean and -1.8 $[-3.4, -0.2]$ for the within-cluster coefficient against -1.3 and +0.5: every cluster then asks the same question, which of two systems is better, and a panel is ranked by one system pair. So at eight clusters and four systems the fall of Section 6 is not a product of drawing the candidates cluster by cluster; at half the clusters the same-systems fall from every system to four is +1.2 $[-0.2, +2.6]$, other numbers of systems and the fixed-label cells below were not run this way, and the two-candidate cells of Tables 19, 24 and 27 are random pairs of systems, which the two-system row shows to understate, at half the clusters, what keeping the same two systems everywhere costs; at eight clusters the two rows do not differ by more than their intervals.

Other seeds. The frozen run reads the main run's 300 splits and 20 subsamples of its own, and its intervals are conditional on those subsamples. Table 26, read after the fact, redraws both under other seeds, point estimates only (`candidate_seeds.py`). On the primary configurations the fall of the grouped mean's gain from all candidates to four at eight clusters, +2.1 under the frozen seed, is +2.1 to +2.4 under four others, the within-cluster coefficient above it at four +1.3 to +1.7 against +1.5, and the two levels without a prediction move more: the grouped mean at four is -1.4 to -1.3 against -1.0 and the within-cluster coefficient -0.0 to +0.4 against +0.5; with every candidate the grouped mean's eight-cluster gain over pooling, +1.2 in the main run, is +0.8 to +1.1. On PPE the grouped mean at four responses, whose predicted negative sign was contradicted, is positive under every seed (+0.7 to +1.3 against +0.9) and the fall is +0.2 to +0.7 against +0.6. On PRM800K a 300-split average at eight problems moves by more than the contrast it is read for: the grouped mean's loss to pooling with six solutions is -7.3 under the frozen seed and -6.1 to -4.3 under twelve others (standard deviation over the seeds 0.8, Monte Carlo standard error from the draws 0.69), inside the interval of the main text and at its upper end, and the fall from six solutions to three, +0.5 under the frozen seed, is +0.6 to +2.2. The sign of every graded contrast on these three sets is the same under every seed but that of C3, the within-cluster coefficient above the grouped mean at four responses on PPE, which is -0.0 to +0.2 against +0.2; the sizes on the small panels at eight units are good to about a point. That the PRM800K spread is the draw and not the script: `candidate_seeds.py` run with the frozen seed reproduces the frozen run's 200 values at both levels, both budgets, both payoffs and every rule exactly (largest difference 0; `results/candidate_seeds_prm800k_frozenreplay_compare_r16.json`).

A fixed number of labels. Tables 27 to 29, read after the fact: the frozen script's bootstrap rerun at the budgets 4 to 32 (`candidate_alloc.py`; at eight clusters and at half its replicates are the frozen run's to the last digit), so that cells $b \times m$ with the same number of labels share their resampled corpora and their splits. On the exploratory configurations the utility of pooling's evaluator does not depend on the split: between the cell with the most candidates a cluster and every other cell of the same N it moves by at most half a point with an interval crossing zero at every N from 24 to 96. The within-cluster Pearson coefficient's does not either down to three candidates (+0.1 $[-0.5, +0.7]$ from four clusters of six to eight of three, +0.2 $[-0.3, +0.8]$ from six of eight to sixteen of three) and loses 1.2 to 1.4 points with two, intervals clear of zero. The grouped mean's falls where the labels are few: +1.0 $[+0.4, +1.5]$ from eight candidates to four at 32 labels, +0.6 $[+0.1, +1.2]$ at 48, +0.3 $[-0.2, +0.9]$ at 64 and +0.3 $[-0.2, +0.9]$ at 96, so in the grouped mean's own utility the cost of four candidates is not clear of zero from sixteen clusters on; +1.2 $[+0.6, +1.9]$ from six to three at 24 labels, and about three points with two at every count. From eight candidates to six it does not fall. FRANK's clusters hold five candidates at the median, so its cells with more than five candidates hold fewer labels than their heading. Against pooling, which is the quantity the protocol grades, the same four contrasts are +0.7 $[-0.0, +1.4]$, +0.6 $[-0.0, +1.2]$, +0.3 $[-0.3, +1.0]$ and +0.7 $[+0.0, +1.4]$: about half a point at every label count, with no trend as clusters are added and one interval of four clear of zero, by less than a twentieth of a point, since pooling's own contrast drifts the other way at the larger counts. The arm identifies an allocation, clusters moving in place of labels; it says that at equal labels the grouped mean's gain over pooling is about half a point lower with four candidates than with eight, and it has no cell for the step from every candidate to eight, which is more than half of the frozen contrast. PPE differs: there no rule pays for more prompts of fewer responses down to

four, the contrasts within half a point of zero up to 48 labels and in favour of more prompts at 64 for pooling and the within-cluster coefficient (-0.8 $[-1.6, -0.1]$ for pooling, -0.8 $[-1.4, -0.2]$ for the within-cluster coefficient and -0.4 $[-1.0, +0.2]$ for the grouped mean at 64 labels, sixteen responses against four), the difference between the within-cluster coefficient and pooling not moving by more than its interval, the grouped mean's own contrast clear of zero in one cell of six; with two responses the grouped mean loses at 32 labels (+0.6 $[+0.1, +1.2]$) and the within-cluster coefficient as well at 24. A practitioner with a number of labels to spend is therefore not told by these panels to label many candidates in few clusters: on the exploratory configurations the split costs the grouped mean about half a point against pooling, and on PPE more prompts are worth as much or more. On WMT23, run the same way, the contrast against pooling is +0.5 $[-1.1, +2.1]$, +0.3 $[-1.3, +1.8]$, +0.7 $[-0.7, +2.0]$ and +1.4 $[+0.1, +2.8]$ at 32, 48, 64 and 96 labels, so there the cost of four candidates grows with the labels and clears zero at 96, and two candidates cost the grouped mean +2.2 points of its own utility at 32. The arm was run with random subsets of candidates and not with the same systems kept in every cluster. No cell of the arm holds more than eight candidates, and the step from every candidate to eight is more than half of the frozen fall at eight clusters, so it was matched in labels separately (`candidate_alloc_all8_r16.py`, the frozen script's subsampling and bootstrap, 128 labels): on the 9 exploratory configurations whose clusters hold sixteen or more candidates, sixteen clusters of eight against eight of sixteen change the grouped mean's gain over pooling by -0.8 $[-1.5, -0.0]$ points, the within-cluster coefficient's by -0.6 $[-1.2, -0.0]$, the per-cluster Pearson mean's by -0.5 $[-1.2, +0.2]$ and the pairwise accuracy's by -0.9 $[-1.6, -0.2]$, while pooling's own utility gains +0.5 $[-0.2, +1.2]$ from the extra clusters; on the 4 PPE tasks with sixteen responses to every prompt the grouped mean's changes by -0.5 $[-1.3, +0.2]$. At equal labels, then, fewer candidates in more clusters cost every within-cluster rule relative to pooling on these configurations, the grouped mean about as much as the others.

Coverage of these intervals. Table 18, read after the fact and finished after the third reading: the manipulation was run inside the resampling world of Appendix AA on the 19 exploratory configurations, 200 corpora each, the analyst subsampling four candidates a cluster, estimating and bootstrapping as the frozen script does. At eight labelled clusters the interval of the graded contrast C5 covers its truth on 98% of corpora and C6's on 97%, the four levels on 92 to 98%, with bootstrap standard deviations 0.93 to 1.20 times the spread of the estimates; at half the levels' intervals cover on 89 to 94% and the two contrasts' on 95% and 93%. So on the exploratory configurations the intervals of the two graded contrasts, which are at eight clusters, are not too narrow; the levels at four candidates are somewhat too narrow at half (91% and 89%), where no hypothesis was stated. The world is a resampling of each corpus and says nothing about PPE, whose intervals are percentiles over its fixed evaluation set, about the addendum's two panels or about WMT23.

The addendum. `prereg/PROTOCOL_candidates_addendum.md`, frozen at 2026-09-19 00:33:18 UTC by this machine with an RFC 3161 token of 00:32:55 by the authority's clock, before the script had computed anything on a subsampled cluster of either panel: the protocol's sentence that PRM800K and RewardBench 2 leave no room to move m is true upward and false downward, and they are the two panels the account was written for. Four graded readings at eight labelled units, the few level half the candidates, no correction, outside the calibration (Table 30). The addendum states that its predictions are 'informed by three sets on which the dose was seen', and two of the four, A2 and A4, grade a level difference that with every candidate kept was already +3.3 $[+2.0, +4.6]$ on RewardBench 2 and +7.3 $[+5.1, +9.6]$ on PRM800K, computable before the freeze; the readings that carried risk are A1, supported, and A3, consistent. On RewardBench 2 the grouped mean's loss to pooling at eight prompts deepens from 2.4 points with the four completions to 5.0 with three and 5.1 with two (A1, +2.8 $[+1.2, +4.4]$, supported; positive on 3 subsets of five and negative on 2, +0.7 $[-1.1, +2.5]$ without Focus, its interval crossing zero on 2 of the five leave-one-outs); the within-cluster coefficient, +0.9 with four, is level with three (+0.0) and -2.1 $[-3.2, -1.0]$ with two, +3.0 $[+1.8, +4.2]$ above the grouped mean (A2, supported); and the per-cluster Pearson mean goes from +0.6 to -1.6 with three. On PRM800K the grouped mean is -7.8 against pooling with three solutions and -7.3 with six under the frozen seed (A3, +0.5 $[-1.6, +2.6]$, consistent); the every-candidate level is the frozen seed's low draw (-6.1 to -4.3 under twelve other seeds, where the fall is +0.6 to +2.2; Table 26), so the grade stands as graded and no floor at eight problems is claimed; at half the problems it goes from -0.2 to -3.3 $[-6.2, -0.4]$ (-3.5 to -1.5 under the other seeds); the within-cluster coefficient is +7.0 $[+5.0, +9.0]$ above it at three (A4, supported) and loses 2.6 $[-4.2, -1.0]$ to pooling with two. The within-cluster coefficient's own gain falls as fast as the

Table 18: Coverage of the manipulation's intervals, measured as Appendix AA measures the main run's and read after the fact (`candidate_calibration.py`): a configuration's units are the population, a corpus is drawn from them with replacement, and inside it the analyst subsamples four candidates a cluster, estimates on 300 splits and bootstraps as the frozen script does. Truth: the mean over corpora, in points of best-of-n recovery; cov.: the share of replications whose estimate ± 1.96 bootstrap standard deviations covers it, and pct. the percentile interval's; ratio: the mean bootstrap standard deviation over the standard deviation of the estimates.

	eight labelled clusters				half the clusters			
	truth	cov. %	pct. %	ratio	truth	cov. %	pct. %	ratio
mean over 19 configurations; 200 replications								
$D_{\mathrm{grp}}(\mathrm{all})$	+1.0	98	92	1.02	+1.4	93	98	0.96
$D_W(\mathrm{all})$	+1.6	94	94	0.93	+1.4	94	98	0.97
$D_{\mathrm{grp}}(4)$	−1.3	96	91	1.03	+0.1	91	95	0.91
$D_W(4)$	+0.1	92	95	0.95	+1.0	89	94	0.83
C5, grp's fall	+2.3	98	100	1.12	+1.2	95	98	1.11
C6, $D_W(4) - D_{\mathrm{grp}}(4)$	+1.4	97	96	1.05	+0.9	93	93	0.89
r_W's fall	+1.5	97	100	1.07	+0.3	98	98	1.23
grp's fall less r_W's	+0.8	98	100	1.20	+0.9	96	99	1.10
mean over 16 configurations, without WMT22; 200 replications								
$D_{\mathrm{grp}}(\mathrm{all})$	+0.9	96	93	1.02	+1.6	94	98	0.95
$D_W(\mathrm{all})$	+1.6	94	92	0.93	+1.5	94	97	0.95
$D_{\mathrm{grp}}(4)$	−1.7	98	92	1.02	+0.2	92	94	0.90
$D_W(4)$	−0.0	94	94	0.93	+1.1	90	94	0.81
C5, grp's fall	+2.5	98	99	1.14	+1.4	95	97	1.10
C6, $D_W(4) - D_{\mathrm{grp}}(4)$	+1.7	97	96	1.03	+0.9	92	92	0.89
r_W's fall	+1.6	97	100	1.09	+0.4	98	98	1.22
grp's fall less r_W's	+1.0	99	100	1.21	+1.0	97	98	1.06

Table 19: The candidates in a labelled cluster, moved under a protocol frozen before any selection rule or utility was computed on a subsampled cluster of real data (Section 6); the clusters, the draws and the scoring are held, the labelled items and the clusters that ask a question move with m. Every labelled cluster keeps m of its candidates, drawn at random; every rule is computed on those, on the draws of Table 5 and Table 15, and the evaluator it selects is scored as there, on clusters no rule used with all their candidates; *all* is the main run, sixteen responses on PPE. Entries: the paired difference against the pooled coefficient computed on the same subsampled clusters, in points, with the estimate ± 1.96 standard deviations of 200 bootstraps of the pool and the evaluation set together (PPE) and of 200 corpus bootstraps of 300 splits (primary), none of them run through the calibration of Appendix AA. A cluster whose m sampled labels are all equal asks no question of any within-cluster rule.

candidates kept, m	2	3	4	6	8	all
PPE, five sets, eight labelled prompts: points of best-of-16 accuracy on the fixed evaluation set						
prompts that ask a question	2.9	4.3	5.2	6.3	6.8	7.7
$r_W - r_T$	+0.3 [−0.4, +1.0]	+0.8 [+0.0, +1.5]	+1.1 [+0.3, +1.8]	+1.3 [+0.5, +2.1]	+1.4 [+0.6, +2.2]	+1.6 [+0.7, +2.5]
grp $- r_T$	−0.1 [−0.8, +0.7]	+0.5 [−0.3, +1.3]	+0.9 [+0.1, +1.7]	+1.0 [+0.2, +1.9]	+1.3 [+0.4, +2.2]	+1.5 [+0.5, +2.5]
per-cluster Pearson mean $- r_T$	−0.1 [−0.8, +0.7]	+0.8 [+0.0, +1.7]	+1.2 [+0.4, +2.0]	+1.3 [+0.4, +2.1]	+1.4 [+0.5, +2.3]	+1.6 [+0.6, +2.5]
Exploratory, 19 configurations, eight labelled clusters: points of best-of-n recovery on the evaluation half						
clusters that ask a question	5.3	6.5	6.9	7.3	7.5	7.7
$r_W - r_T$	−1.6 [−2.4, −0.8]	−0.3 [−1.1, +0.5]	+0.5 [−0.3, +1.3]	+0.9 [+0.1, +1.8]	+1.0 [+0.1, +2.0]	+1.6 [+0.7, +2.6]
grp $- r_T$	−4.4 [−5.4, −3.3]	−2.2 [−3.2, −1.3]	−1.0 [−1.9, −0.0]	−0.1 [−1.0, +0.8]	+0.0 [−0.9, +1.0]	+1.2 [+0.1, +2.2]
per-cluster Pearson mean $- r_T$	−4.4 [−5.5, −3.3]	−1.5 [−2.4, −0.5]	−0.0 [−0.9, +0.9]	+0.5 [−0.4, +1.4]	+0.9 [−0.1, +1.8]	+1.7 [+0.6, +2.8]

grouped mean's on both panels, the difference in falls -0.3 $[-1.8, +1.2]$ and -0.3 $[-2.4, +1.7]$, so A2 and A4 do not show it holding up better. So on RewardBench 2 at eight prompts removing completions costs the grouped mean and the within-cluster coefficient against pooling, on two subsets of five; on PRM800K at eight problems the grouped mean's fall is not clear of zero. That the four and six candidates these panels hold are what puts the grouped mean below pooling there is not shown by this: no panel can be given more candidates than it has.

O Ordering, Naming and Gain by Configuration

`scripts/decision/separation.py` reads the targets files and tabulates for every configuration (Table 32) the grouped mean against pooling at half the clusters against the out-of-sample target,

Table 20: The six graded contrasts of the candidate-count protocol at eight labelled units, in points (best-of-16 accuracy on PPE, best-of-n recovery on the exploratory configurations). $D_r(m)$ is the paired difference between the evaluator rule r selects and the one the pooled coefficient selects when every labelled cluster keeps m candidates. Six graded readings, PPE the primary test and the exploratory configurations the replication, no primary hypothesis among them and no correction; 200 and 200 bootstrap replicates. The grade reads the estimate ± 1.96 bootstrap standard deviations, as the protocol fixed; the percentile interval of the same bootstrap and, on PPE, of the bootstrap of the evaluation set alone, which is conditional on the pool, are beside it.

	Set	Prediction	Estimate ± 1.96 SD	Percentile	Evaluation set only	Grade
C1	PPE	$D_{\mathrm{grp}}(16) - D_{\mathrm{grp}}(4) > 0$	+0.6 [−0.1, +1.2]	[−0.1, +1.2]	[+0.3, +0.9]	cons.
C2	PPE	$D_{\mathrm{grp}}(4) < 0$	+0.9 [+0.1, +1.7]	[+0.0, +1.5]	[+0.4, +1.4]	contr.
C3	PPE	$D_W(4) - D_{\mathrm{grp}}(4) > 0$	+0.2 [−0.3, +0.6]	[−0.4, +0.5]	[−0.1, +0.4]	cons.
C4	PPE	$D_W(4) > 0$	+1.1 [+0.3, +1.8]	[+0.3, +1.6]	[+0.6, +1.5]	supp.
C5	Exploratory	$D_{\mathrm{grp}}(\mathrm{all}) - D_{\mathrm{grp}}(4) > 0$	+2.1 [+1.2, +3.0]	[+1.4, +3.2]	–	supp.
C6	Exploratory	$D_W(4) - D_{\mathrm{grp}}(4) > 0$	+1.5 [+0.8, +2.1]	[+0.9, +2.2]	–	supp.

Table 21: PPE by task at eight labelled prompts: the paired difference in points of best-of-16 accuracy on the fixed evaluation set between the evaluator the grouped mean (grp) or the within-cluster coefficient (r_W) selects and the one the pooled coefficient selects, with sixteen, four and two responses a prompt; the estimate ± 1.96 standard deviations of 200 bootstraps of the pool and the evaluation set together, each task resampled on its own. One task, MBPP-Plus, supplies 80% of the five-task mean at sixteen responses and the positive sign at four. Below: the protocol's four PPE contrasts with each task left out in turn; the grades are the frozen rule applied to the four remaining tasks, read after the fact. Estimates average 500 selection draws and 20 subsamples of responses; a bootstrap replicate averages 300 draws, and the intervals are conditional on the subsamples.

	grp $- r_T$, responses a prompt			$r_W - r_T$, responses a prompt		
Task	16	4	2	16	4	2
MMLU-Pro	−0.7 [−2.5, +1.1]	−1.0 [−2.4, +0.5]	−1.2 [−2.5, +0.2]	−0.1 [−1.3, +1.1]	−0.2 [−1.2, +0.9]	−0.8 [−1.9, +0.3]
MATH	+0.0 [−1.6, +1.6]	−1.5 [−2.8, −0.1]	−1.7 [−3.2, −0.1]	+0.4 [−0.8, +1.6]	−0.6 [−1.8, +0.5]	−0.8 [−2.1, +0.6]
GPQA	+0.5 [−1.1, +2.1]	+0.3 [−1.0, +1.6]	−0.2 [−1.4, +0.9]	+0.4 [−0.9, +1.7]	+0.1 [−1.0, +1.2]	−0.5 [−1.7, +0.6]
IFEval	+1.6 [−0.4, +3.6]	+1.1 [−0.5, +2.7]	+0.7 [−0.9, +2.2]	+2.0 [−0.2, +4.2]	+1.0 [−0.6, +2.6]	+0.6 [−0.9, +2.1]
MBPP-Plus	+5.9 [+2.0, +9.7]	+5.5 [+2.4, +8.7]	+2.0 [−0.7, +4.7]	+5.2 [+1.9, +8.5]	+5.0 [+2.2, +7.8]	+3.1 [+0.5, +5.7]
five tasks	+1.5 [+0.5, +2.5]	+0.9 [+0.1, +1.7]		+1.6 [+0.7, +2.5]	+1.1 [+0.3, +1.8]	
without MBPP-Plus	+0.4 [−0.5, +1.2]	−0.3 [−1.0, +0.4]		+0.7 [−0.1, +1.4]	+0.1 [−0.5, +0.7]	

The four graded contrasts leaving out one task: estimate, interval, the grade it would get
without none: C1 +0.6 [−0.1, +1.2] cons.; C2 +0.9 [+0.1, +1.7] contr.; C3 +0.2 [−0.3, +0.6] cons.; C4 +1.1 [+0.3, +1.8] supp.
without MMLU-Pro: C1 +0.6 [−0.2, +1.5] cons.; C2 +1.4 [+0.4, +2.3] contr.; C3 −0.0 [−0.5, +0.5] not supp.; C4 +1.4 [+0.5, +2.2] supp.
without MATH: C1 +0.3 [−0.4, +1.1] cons.; C2 +1.5 [+0.5, +2.5] contr.; C3 −0.0 [−0.5, +0.5] not supp.; C4 +1.5 [+0.6, +2.4] supp.
without GPQA: C1 +0.6 [−0.2, +1.5] cons.; C2 +1.0 [+0.1, +2.0] contr.; C3 +0.2 [−0.3, +0.8] cons.; C4 +1.3 [+0.5, +2.1] supp.
without IFEval: C1 +0.6 [−0.2, +1.3] cons.; C2 +0.9 [−0.0, +1.7] not supp.; C3 +0.2 [−0.3, +0.8] cons.; C4 +1.1 [+0.2, +1.9] supp.
without MBPP-Plus: C1 +0.6 [+0.0, +1.2] supp.; C2 −0.3 [−1.0, +0.4] cons.; C3 +0.3 [−0.1, +0.8] cons.; C4 +0.1 [−0.5, +0.7] cons.

the complement on the released panels and PRM800K and the fixed measurement set on PPE: the share of draws on which the two select different evaluators, the utility difference of their selected evaluators, which that share times the difference given disagreement gives exactly, the share of disagreeing draws the grouped mean's evaluator wins, and for each rule the share of draws selecting the target's best evaluator and the regret share given a miss; beside them the two properties read off the whole corpus. Under best of n the regret-share gap of +5.4 points splits exactly into a naming part, the difference in miss rates weighted by the mean loss given a miss, of +4.3 and a loss part of +1.1. Across the 26 configurations the utility difference correlates with the disagreement rate at a Spearman +0.31 and with the difference given disagreement at +0.89; the error gap with the regret gap at +0.56; the regret-share gap with the gap of the two rules' limit regrets at +0.56 and with the margin at +0.11. By pooling's limit choice, the configurations where it is the best evaluator gain −0.2 points of the span on average, from −41 to +40 by configuration, and the others +11.0, from −17 to +57. Under the pairwise choice the utility difference is +0.7 points, the grouped mean wins 69% of disagreeing draws, and the limit groups, 17 and 9 configurations, gain +1.0 and +18.1 points of the span. At eight clusters under best of n the difference is +1.0 points, the grouped mean wins 53% of disagreeing draws, and the limit groups gain +4.4 and +3.8. The configurations are dependent units and the split reads the corpus that also pays, so these describe where the gain sat on these panels and are not a test.

Table 22: Beside every set mean of the candidate-count manipulation on the released panels, at eight labelled clusters, in points of best-of-n recovery: the estimate ± 1.96 corpus-bootstrap standard deviations with the percentile interval of the same replicates in parentheses; the configurations on which the quantity is positive and negative; the configuration with the largest value and the mean without it; and on how many of the leave-one-configuration-out means the interval crosses zero. C5 and C6 are graded; the others are levels, of which $D_{\mathrm{grp}}(4)$ and $D_W(4)$ are the two the protocol said carried no prediction. WMT23 was run and not graded. Estimates average 300 splits and 20 subsamples of candidates; intervals are over 200 corpus-bootstrap replicates of 300 splits each, conditional on those subsamples, and were not run through the calibration.

Quantity	Estimate	+ / −	Largest configuration	Without it	Crossing zero
Exploratory, 19 configurations					
C5, $D_{\mathrm{grp}}(\mathrm{all}) - D_{\mathrm{grp}}(4)$	+2.1 [+1.2, +3.0] (+1.4, +3.2)	16 / 3	SummEval consistency (+6.4)	+1.9 [+1.0, +2.8]	0 of 19
C6, $D_W(4) - D_{\mathrm{grp}}(4)$	+1.5 [+0.8, +2.1] (+0.9, +2.2)	15 / 4	SummEval fluency (+6.8)	+1.2 [+0.5, +1.8]	0 of 19
$D_{\mathrm{grp}}(4)$	−1.0 [−1.9, −0.0] (−2.4, −0.7)	8 / 11	SummEval fluency (−7.8)	−0.6 [−1.5, +0.4]	9 of 19
$D_W(4)$	+0.5 [−0.3, +1.3] (−0.8, +0.7)	8 / 11	HANNA relevance (+9.9)	−0.0 [−0.8, +0.7]	19 of 19
$D_{\mathrm{grp}}(6)$	−0.1 [−1.0, +0.8] (−1.4, +0.3)	8 / 11	HANNA relevance (+8.7)	−0.6 [−1.5, +0.3]	19 of 19
$D_W(6)$	+0.9 [+0.1, +1.8] (−0.3, +1.5)	9 / 9	HANNA relevance (+13.7)	+0.2 [−0.5, +1.0]	3 of 19
$D_{\mathrm{grp}}(\mathrm{all})$	+1.2 [+0.1, +2.2] (−0.3, +1.7)	11 / 7	HANNA relevance (+10.5)	+0.6 [−0.3, +1.6]	3 of 19
$D_W(\mathrm{all})$	+1.6 [+0.7, +2.6] (+0.5, +2.4)	13 / 6	HANNA relevance (+14.9)	+0.9 [+0.1, +1.8]	0 of 19
WMT23 MQM, 3 configurations					
C5, $D_{\mathrm{grp}}(\mathrm{all}) - D_{\mathrm{grp}}(4)$	+1.8 [+0.4, +3.2] (+1.1, +3.8)	3 / 0	WMT23 zhen (+3.7)	+0.8 [−0.8, +2.5]	1 of 3
C6, $D_W(4) - D_{\mathrm{grp}}(4)$	+0.7 [−0.5, +2.0] (+0.4, +2.8)	2 / 1	WMT23 ende (+2.0)	+0.1 [−1.3, +1.6]	2 of 3
$D_{\mathrm{grp}}(4)$	−1.8 [−3.2, −0.4] (−3.9, −1.0)	0 / 3	WMT23 zhen (−3.3)	−1.1 [−3.0, +0.8]	2 of 3
$D_W(4)$	−1.1 [−2.2, +0.1] (−2.1, +0.2)	0 / 3	WMT23 zhen (−2.1)	−0.6 [−2.1, +0.9]	2 of 3
$D_{\mathrm{grp}}(6)$	−1.0 [−2.3, +0.3] (−2.7, +0.0)	0 / 3	WMT23 zhen (−1.7)	−0.7 [−2.5, +1.0]	3 of 3
$D_W(6)$	−0.6 [−1.6, +0.4] (−1.5, +0.7)	2 / 1	WMT23 zhen (−2.3)	+0.2 [−1.1, +1.5]	3 of 3
$D_{\mathrm{grp}}(\mathrm{all})$	−0.0 [−1.4, +1.3] (−1.5, +1.2)	1 / 1	WMT23 ende (−0.5)	+0.2 [−1.3, +1.7]	3 of 3
$D_W(\mathrm{all})$	−0.1 [−1.1, +0.9] (−0.5, +1.4)	2 / 1	WMT23 zhen (−1.0)	+0.4 [−0.8, +1.5]	3 of 3

Table 23: What else moves with the candidates kept, at eight labelled clusters: the mean number of the eight clusters that still ask a question of a within-cluster rule (a cluster whose kept labels are all equal asks none), and the mean number of evaluators tied at the top of each rule's ranking, among whom the draw's priority decides; means over the configurations of a set and the draws. The protocol listed both as reported.

candidates kept, m	2	3	4	6	8	all
PPE						
clusters that ask a question	2.9	4.3	5.2	6.3	6.8	7.7
tied at the top, r_T	1.00	1.00	1.00	1.00	1.00	1.00
tied at the top, r_W	1.14	1.01	1.00	1.00	1.00	1.00
tied at the top, grp	2.60	1.55	1.16	1.01	1.00	1.00
tied at the top, per-cluster Pearson mean	2.60	1.01	1.00	1.00	1.00	1.00
tied at the top, pairwise accuracy	2.60	1.57	1.31	1.10	1.06	1.01
tied at the top, best of n	2.60	2.08	1.89	1.76	1.70	1.62
Exploratory						
clusters that ask a question	5.3	6.5	6.9	7.3	7.5	7.7
tied at the top, r_T	1.04	1.01	1.00	1.00	1.00	1.00
tied at the top, r_W	1.07	1.01	1.00	1.00	1.00	1.00
tied at the top, grp	3.28	1.45	1.08	1.01	1.01	1.01
tied at the top, per-cluster Pearson mean	3.28	1.01	1.00	1.00	1.00	1.00
tied at the top, pairwise accuracy	3.28	1.80	1.39	1.16	1.11	1.05
tied at the top, best of n	2.92	2.25	2.04	1.91	1.88	1.85
WMT23 MQM						
clusters that ask a question	5.6	6.6	7.0	7.4	7.6	7.7
tied at the top, r_T	1.00	1.00	1.00	1.00	1.00	1.00
tied at the top, r_W	1.00	1.00	1.00	1.00	1.00	1.00
tied at the top, grp	5.09	1.52	1.08	1.00	1.00	1.00
tied at the top, per-cluster Pearson mean	5.09	1.00	1.00	1.00	1.00	1.00
tied at the top, pairwise accuracy	5.09	1.99	1.51	1.20	1.11	1.04
tied at the top, best of n	4.36	2.49	2.10	1.80	1.55	1.46

The accounting of the utility difference. The utility difference is an exact accounting of regret against the evaluation half's best evaluator: with h_r the rate at which rule r selects it and L_r the mean regret when it does not, $U_G - U_T = (h_G - h_T)\bar{L} + (1 - \bar{h})(L_T - L_G)$, a naming part and a loss part. That is the symmetric member of a family of exact splits under which a set row's parts move by up to 0.5 points and a configuration's by 2.7, the larger part changing on 10 of 34 configurations, so the rates and the losses are the quantities, the parts are point estimates without an interval, and a part's size is read only where every member agrees. At half the clusters the primary gain is +1.4 naming

Table 24: The candidate-count manipulation in full: every rule against the pooled coefficient computed on the same subsampled clusters, by the number of candidates kept, m. Intervals as in Table 19. WMT23 is reported and was not graded. With two candidates a cluster's Pearson and Spearman coefficients are both ± 1, so the grouped mean, the per-cluster Pearson mean and the pairwise accuracy coincide.

rule $-\ r_T$; candidates kept, m	2	3	4	6	8	all
PPE, 8 prompts: points of best-of-16 accuracy						
r_W	+0.3 [−0.4, +1.0]	+0.8 [+0.0, +1.5]	+1.1 [+0.3, +1.8]	+1.3 [+0.5, +2.1]	+1.4 [+0.6, +2.2]	+1.6 [+0.7, +2.5]
grp	−0.1 [−0.8, +0.7]	+0.5 [−0.3, +1.3]	+0.9 [+0.1, +1.7]	+1.0 [+0.2, +1.9]	+1.3 [+0.4, +2.2]	+1.5 [+0.5, +2.5]
pairwise accuracy	−0.1 [−0.8, +0.7]	+0.6 [−0.2, +1.4]	+0.9 [+0.1, +1.6]	+1.1 [+0.3, +1.9]	+1.3 [+0.4, +2.3]	+1.5 [+0.5, +2.5]
per-cluster Pearson mean	−0.1 [−0.8, +0.7]	+0.8 [+0.0, +1.7]	+1.2 [+0.4, +2.0]	+1.3 [+0.4, +2.1]	+1.4 [+0.5, +2.3]	+1.6 [+0.6, +2.5]
best of n	−0.1 [−0.8, +0.7]	+0.1 [−0.6, +0.9]	+0.2 [−0.6, +1.0]	−0.2 [−1.0, +0.7]	−0.2 [−1.1, +0.8]	−0.1 [−1.2, +1.0]
PPE, 24 prompts: points of best-of-16 accuracy						
r_W	+0.7 [−0.4, +1.8]	+0.9 [−0.3, +2.2]	+1.1 [−0.2, +2.3]	+1.2 [−0.1, +2.6]	+1.1 [−0.3, +2.5]	+1.3 [−0.2, +2.7]
grp	+0.4 [−0.7, +1.5]	+0.4 [−0.9, +1.6]	+0.6 [−0.6, +1.9]	+0.7 [−0.7, +2.1]	+0.6 [−0.8, +2.1]	+0.7 [−0.9, +2.2]
pairwise accuracy	+0.4 [−0.7, +1.5]	+0.4 [−0.8, +1.6]	+0.6 [−0.7, +1.9]	+0.7 [−0.7, +2.1]	+0.7 [−0.8, +2.1]	+0.7 [−0.9, +2.3]
per-cluster Pearson mean	+0.4 [−0.7, +1.5]	+0.7 [−0.6, +1.9]	+0.8 [−0.5, +2.1]	+0.9 [−0.5, +2.3]	+0.9 [−0.5, +2.4]	+0.9 [−0.6, +2.4]
best of n	+0.4 [−0.7, +1.5]	+0.1 [−1.1, +1.4]	+0.3 [−1.0, +1.5]	−0.1 [−1.5, +1.2]	−0.2 [−1.7, +1.3]	−0.5 [−2.2, +1.2]
PPE, 48 prompts: points of best-of-16 accuracy						
r_W	+0.5 [−0.7, +1.8]	+0.7 [−0.6, +2.0]	+0.6 [−0.8, +1.9]	+0.7 [−0.8, +2.2]	+0.8 [−0.7, +2.4]	+0.9 [−0.7, +2.5]
grp	+0.2 [−1.1, +1.5]	+0.0 [−1.3, +1.4]	−0.1 [−1.5, +1.3]	+0.0 [−1.6, +1.6]	+0.0 [−1.5, +1.6]	+0.1 [−1.6, +1.8]
pairwise accuracy	+0.2 [−1.1, +1.5]	+0.1 [−1.3, +1.4]	−0.1 [−1.6, +1.3]	+0.0 [−1.6, +1.6]	+0.1 [−1.5, +1.7]	+0.1 [−1.6, +1.8]
per-cluster Pearson mean	+0.2 [−1.1, +1.5]	+0.2 [−1.1, +1.5]	+0.1 [−1.3, +1.5]	+0.3 [−1.3, +1.8]	+0.4 [−1.2, +1.9]	+0.6 [−1.0, +2.3]
best of n	+0.2 [−1.1, +1.5]	−0.2 [−1.6, +1.1]	−0.6 [−2.0, +0.8]	−0.6 [−2.2, +0.9]	−0.7 [−2.3, +0.9]	−0.9 [−2.8, +0.9]
Exploratory, 8 clusters: points of best-of-n recovery						
r_W	−1.6 [−2.4, −0.8]	−0.3 [−1.1, +0.5]	+0.5 [−0.3, +1.3]	+0.9 [+0.1, +1.8]	+1.0 [+0.1, +2.0]	+1.6 [+0.7, +2.6]
grp	−4.4 [−5.4, −3.3]	−2.2 [−3.2, −1.3]	−1.0 [−1.9, −0.0]	−0.1 [−1.0, +0.8]	+0.0 [−0.9, +1.0]	+1.2 [+0.1, +2.2]
pairwise accuracy	−4.4 [−5.4, −3.3]	−2.3 [−3.3, −1.3]	−0.6 [−1.6, +0.3]	+0.3 [−0.6, +1.3]	+0.6 [−0.4, +1.6]	+1.7 [+0.7, +2.8]
per-cluster Pearson mean	−4.4 [−5.5, −3.3]	−1.5 [−2.4, −0.5]	−0.0 [−0.9, +0.9]	+0.5 [−0.4, +1.4]	+0.9 [−0.1, +1.8]	+1.7 [+0.6, +2.8]
best of n	−3.8 [−4.8, −2.7]	−3.3 [−4.3, −2.2]	−2.9 [−4.0, −1.8]	−3.6 [−4.9, −2.4]	−3.8 [−5.2, −2.5]	−4.4 [−6.0, −2.9]
Exploratory, 8 clusters: points of pairwise accuracy						
r_W	−0.8 [−1.0, −0.6]	−0.3 [−0.5, −0.1]	+0.1 [−0.2, +0.3]	+0.3 [+0.1, +0.5]	+0.4 [+0.2, +0.7]	+0.7 [+0.4, +0.9]
grp	−1.9 [−2.1, −1.6]	−1.1 [−1.4, −0.9]	−0.6 [−0.8, −0.3]	−0.2 [−0.5, +0.1]	−0.0 [−0.3, +0.3]	+0.5 [+0.2, +0.8]
pairwise accuracy	−1.9 [−2.1, −1.6]	−1.1 [−1.4, −0.9]	−0.4 [−0.7, −0.1]	−0.0 [−0.3, +0.3]	+0.2 [−0.1, +0.5]	+0.7 [+0.4, +1.0]
per-cluster Pearson mean	−1.9 [−2.1, −1.6]	−0.8 [−1.0, −0.6]	−0.2 [−0.4, +0.1]	+0.1 [−0.1, +0.4]	+0.3 [+0.0, +0.6]	+0.7 [+0.4, +1.0]
best of n	−1.6 [−1.9, −1.3]	−1.7 [−1.9, −1.4]	−1.6 [−1.9, −1.3]	−2.1 [−2.4, −1.7]	−2.1 [−2.4, −1.8]	−2.5 [−2.9, −2.1]
Exploratory, half the clusters: points of best-of-n recovery						
r_W	+0.0 [−1.0, +1.1]	+0.8 [−0.4, +2.0]	+0.6 [−0.7, +1.9]	+1.2 [−0.3, +2.6]	+1.1 [−0.4, +2.7]	+1.2 [−0.5, +2.9]
grp	−1.9 [−3.3, −0.5]	−0.4 [−1.7, +0.9]	−0.3 [−1.6, +1.1]	+1.1 [−0.3, +2.6]	+1.0 [−0.7, +2.7]	+1.6 [−0.2, +3.5]
pairwise accuracy	−1.9 [−3.3, −0.5]	−0.3 [−1.6, +1.1]	−0.1 [−1.5, +1.3]	+1.4 [−0.1, +2.8]	+1.3 [−0.3, +3.0]	+2.1 [+0.3, +4.0]
per-cluster Pearson mean	−1.9 [−3.3, −0.4]	+0.1 [−1.2, +1.3]	+0.3 [−1.1, +1.6]	+1.5 [+0.0, +3.1]	+1.5 [−0.2, +3.3]	+2.0 [+0.1, +4.0]
best of n	−1.3 [−2.5, +0.0]	−0.3 [−1.7, +1.0]	−0.5 [−1.9, +0.9]	+0.1 [−1.5, +1.7]	+0.1 [−1.7, +2.0]	+0.2 [−2.0, +2.3]
Exploratory, half the clusters: points of pairwise accuracy						
r_W	+0.1 [−0.2, +0.5]	+0.4 [+0.1, +0.8]	+0.4 [+0.0, +0.8]	+0.6 [+0.2, +1.0]	+0.6 [+0.2, +1.1]	+0.6 [+0.1, +1.0]
grp	−0.7 [−1.1, −0.2]	−0.0 [−0.4, +0.4]	+0.2 [−0.2, +0.6]	+0.6 [+0.2, +1.1]	+0.8 [+0.3, +1.2]	+0.9 [+0.4, +1.4]
pairwise accuracy	−0.7 [−1.1, −0.2]	+0.0 [−0.4, +0.4]	+0.3 [−0.1, +0.7]	+0.7 [+0.3, +1.1]	+0.9 [+0.4, +1.3]	+1.0 [+0.5, +1.5]
per-cluster Pearson mean	−0.7 [−1.1, −0.2]	+0.2 [−0.2, +0.6]	+0.4 [+0.0, +0.9]	+0.8 [+0.3, +1.2]	+0.9 [+0.4, +1.3]	+0.8 [+0.4, +1.3]
best of n	−0.4 [−0.8, −0.0]	−0.3 [−0.7, +0.2]	−0.4 [−0.9, +0.1]	−0.6 [−1.1, −0.0]	−0.6 [−1.1, −0.0]	−0.6 [−1.3, +0.1]
WMT23 MQM, 8 clusters: points of best-of-n recovery						
r_W	−1.4 [−2.8, −0.0]	−2.3 [−3.6, −1.1]	−1.1 [−2.2, +0.1]	−0.6 [−1.6, +0.4]	+0.3 [−0.6, +1.2]	−0.1 [−1.1, +0.9]
grp	−4.1 [−5.7, −2.4]	−3.7 [−5.2, −2.3]	−1.8 [−3.2, −0.4]	−1.0 [−2.3, +0.3]	−0.6 [−2.0, +0.7]	−0.0 [−1.4, +1.3]
pairwise accuracy	−4.1 [−5.7, −2.4]	−3.6 [−5.0, −2.2]	−2.0 [−3.4, −0.6]	−1.0 [−2.4, +0.3]	−0.6 [−1.9, +0.7]	+0.2 [−1.1, +1.5]
per-cluster Pearson mean	−4.1 [−5.7, −2.4]	−3.6 [−5.1, −2.2]	−1.4 [−2.9, +0.1]	−0.5 [−1.8, +0.8]	+0.2 [−1.0, +1.3]	+0.3 [−0.9, +1.5]
best of n	−4.4 [−6.1, −2.7]	−4.2 [−5.8, −2.6]	−2.7 [−4.2, −1.3]	−4.3 [−5.7, −2.9]	−2.7 [−4.2, −1.3]	−3.2 [−4.7, −1.8]
WMT23 MQM, 8 clusters: points of pairwise accuracy						
r_W	−0.7 [−1.0, −0.4]	−0.8 [−1.1, −0.5]	−0.4 [−0.7, −0.1]	−0.3 [−0.5, −0.0]	−0.1 [−0.3, +0.1]	−0.1 [−0.3, +0.1]
grp	−1.4 [−1.8, −1.0]	−1.3 [−1.6, −0.9]	−0.7 [−1.1, −0.4]	−0.5 [−0.9, −0.2]	−0.3 [−0.6, −0.0]	−0.2 [−0.5, +0.1]
pairwise accuracy	−1.4 [−1.8, −1.0]	−1.2 [−1.6, −0.9]	−0.7 [−1.0, −0.3]	−0.4 [−0.7, −0.0]	−0.1 [−0.4, +0.1]	+0.1 [−0.2, +0.4]
per-cluster Pearson mean	−1.4 [−1.8, −1.0]	−1.1 [−1.5, −0.8]	−0.4 [−0.8, −0.1]	−0.2 [−0.5, +0.1]	+0.0 [−0.2, +0.3]	+0.2 [−0.1, +0.4]
best of n	−1.5 [−1.9, −1.0]	−1.5 [−1.9, −1.1]	−1.1 [−1.5, −0.7]	−1.7 [−2.0, −1.3]	−1.2 [−1.6, −0.9]	−1.4 [−1.7, −1.0]
WMT23 MQM, half the clusters: points of best-of-n recovery						
r_W	+0.2 [−1.4, +1.7]	+0.7 [−1.0, +2.3]	+1.0 [−0.8, +2.7]	+0.9 [−1.0, +2.7]	+1.1 [−0.8, +3.0]	+0.7 [−1.4, +2.8]
grp	−1.1 [−3.5, +1.2]	−0.0 [−2.2, +2.2]	−0.1 [−2.6, +2.5]	−0.2 [−2.7, +2.3]	−0.3 [−2.9, +2.3]	−0.6 [−3.6, +2.4]
pairwise accuracy	−1.1 [−3.5, +1.2]	−0.1 [−2.3, +2.1]	−0.3 [−2.8, +2.2]	−0.1 [−2.6, +2.4]	−0.1 [−2.7, +2.5]	−0.4 [−3.3, +2.5]
per-cluster Pearson mean	−1.1 [−3.5, +1.2]	−0.2 [−2.4, +1.9]	+0.0 [−2.4, +2.4]	−0.2 [−2.5, +2.0]	+0.1 [−2.3, +2.4]	−0.1 [−2.6, +2.3]
best of n	−0.8 [−3.3, +1.7]	+0.4 [−1.9, +2.7]	+0.3 [−2.2, +2.8]	+0.6 [−2.0, +3.2]	+1.0 [−1.8, +3.9]	+1.8 [−2.1, +5.6]
WMT23 MQM, half the clusters: points of pairwise accuracy						
r_W	−0.1 [−0.4, +0.2]	+0.0 [−0.3, +0.4]	+0.2 [−0.2, +0.5]	+0.1 [−0.3, +0.4]	+0.2 [−0.1, +0.6]	+0.1 [−0.3, +0.5]
grp	−0.5 [−0.9, −0.0]	−0.0 [−0.5, +0.4]	−0.1 [−0.6, +0.4]	−0.2 [−0.7, +0.4]	−0.1 [−0.7, +0.5]	−0.3 [−1.0, +0.4]
pairwise accuracy	−0.5 [−0.9, −0.0]	+0.1 [−0.4, +0.5]	+0.1 [−0.4, +0.6]	+0.2 [−0.3, +0.7]	+0.4 [−0.2, +0.9]	+0.3 [−0.4, +0.9]
per-cluster Pearson mean	−0.5 [−0.9, −0.0]	+0.2 [−0.3, +0.6]	+0.3 [−0.2, +0.7]	+0.3 [−0.2, +0.8]	+0.5 [−0.0, +1.0]	+0.5 [−0.1, +1.0]
best of n	−0.5 [−0.9, −0.0]	−0.1 [−0.6, +0.3]	−0.2 [−0.7, +0.3]	−0.3 [−0.8, +0.2]	−0.2 [−0.7, +0.3]	−0.6 [−1.2, −0.0]

Table 25: The manipulation with the same systems kept in every labelled cluster, read after the fact, beside the frozen random subsets on the same 15 exploratory configurations, those whose candidates carry a system label (the four speechocean762 configurations, whose clusters are speakers, do not). Paired difference in points of best-of-n recovery between the evaluator a rule selects and the one the pooled coefficient selects on the same kept candidates, at eight labelled clusters and at half; 20 draws of the m systems a configuration; estimate ± 1.96 corpus-bootstrap standard deviations of the mean over configurations, the frozen script's streams. Beside the grouped mean's fall: configurations on which it is positive and negative. Estimates average 300 splits and 20 subsamples of candidates; intervals are over 200 corpus-bootstrap replicates of 300 splits each, conditional on those subsamples, and were not run through the calibration.

	eight labelled clusters			half the clusters		
	$r_W - r_T$	grp $- r_T$	per-cluster Pearson $- r_T$	$r_W - r_T$	grp $- r_T$	per-cluster Pearson $- r_T$
the same m systems in every labelled cluster						
$m = 2$	-1.7 [-2.7, -0.7]	-3.7 [-4.9, -2.4]	-3.7 [-4.9, -2.4]	-1.8 [-3.4, -0.2]	-4.6 [-6.4, -2.8]	-4.6 [-6.4, -2.8]
$m = 4$	+0.4 [-0.5, +1.3]	-0.9 [-1.9, +0.0]	-0.1 [-0.9, +0.8]	+1.1 [-0.4, +2.6]	+0.3 [-1.1, +1.8]	+1.0 [-0.5, +2.5]
$m = 8$	+1.4 [+0.3, +2.5]	+0.6 [-0.4, +1.6]	+1.3 [+0.2, +2.4]	+1.7 [-0.2, +3.6]	+1.7 [-0.2, +3.5]	+1.9 [-0.1, +3.9]
all	+1.9 [+0.8, +3.1]	+1.5 [+0.3, +2.6]	+2.2 [+0.9, +3.4]	+1.7 [-0.4, +3.8]	+1.5 [-0.6, +3.7]	+2.0 [-0.2, +4.2]
fall, all to 4	+1.5 [+0.8, +2.3]	+2.4 [+1.4, +3.4] 13/2	+2.2 [+1.2, +3.2]	+0.6 [-0.6, +1.8]	+1.2 [-0.2, +2.6] 12/3	+1.0 [-0.4, +2.4]
$D_W(4) - D_{\mathrm{grp}}(4)$		+1.3 [+0.6, +2.0]			+0.8 [-0.1, +1.6]	
grp's fall less r_W's		+0.9 [+0.0, +1.7]			+0.6 [-0.7, +1.9]	
a random subset of m candidates in every cluster (frozen)						
$m = 2$	-1.4 [-2.2, -0.6]	-4.5 [-5.7, -3.2]	-4.5 [-5.8, -3.3]	+0.5 [-0.8, +1.8]	-1.3 [-3.0, +0.4]	-1.3 [-2.9, +0.4]
$m = 4$	+0.9 [+0.0, +1.8]	-0.6 [-1.6, +0.4]	+0.5 [-0.4, +1.4]	+1.0 [-0.6, +2.6]	+0.1 [-1.5, +1.6]	+0.7 [-0.9, +2.3]
$m = 8$	+1.5 [+0.4, +2.6]	+0.5 [-0.6, +1.5]	+1.5 [+0.4, +2.6]	+1.6 [-0.4, +3.6]	+1.2 [-0.8, +3.1]	+1.7 [-0.3, +3.7]
all	+1.9 [+0.8, +3.1]	+1.5 [+0.3, +2.6]	+2.2 [+0.9, +3.4]	+1.7 [-0.4, +3.8]	+1.5 [-0.6, +3.7]	+2.0 [-0.2, +4.2]
fall, all to 4	+1.0 [+0.4, +1.7]	+2.1 [+1.1, +3.1] 12/3	+1.7 [+0.7, +2.6]	+0.7 [-0.3, +1.6]	+1.5 [+0.2, +2.7] 12/3	+1.3 [+0.1, +2.5]
$D_W(4) - D_{\mathrm{grp}}(4)$		+1.5 [+0.7, +2.3]			+0.9 [+0.2, +1.7]	
grp's fall less r_W's		+1.1 [+0.3, +1.8]			+0.8 [-0.3, +1.9]	

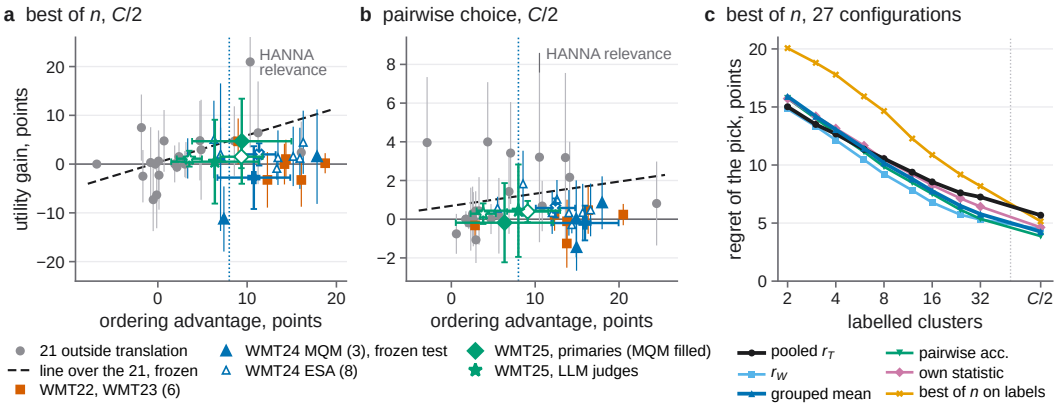


Figure 8: **Ordering against choice by configuration, the main text's Figure 1b before averaging.** (a, b) One point per configuration at half its clusters ($C/2$): the grouped mean's ordering advantage over pooling (points of selection error over evaluator pairs) against the utility gain of the evaluator it selects (± 1.96 bootstrap standard deviations), under best of n and under the pairwise choice. Grey, the 21 configurations outside machine translation and (dashed) their line, frozen before WMT24 was read; vermilion, WMT22 and WMT23; blue, WMT24 (filled MQM, hollow ESA) with each set's mean and interval; green, the WMT25 set means of the fourth protocol, diamonds for the year's primary metrics and stars for thirteen LLM judges (filled MQM, hollow ESA); dotted, M1's threshold. (c) Each rule's regret on the evaluation half by the clusters it reads, best of n, mean over the 27 configurations read first (after the third protocol was graded).

and +0.2 loss, the naming part HANNA relevance, both parts within a point of zero without it under every member; per pair the ordering advantage is +3.0 points on the pairs with the best evaluator, the displacement of that evaluator's mean rank, and +5.4 on the rest.

Table 26: The manipulation's point estimates under other seeds, read after the fact: the splits and the subsamples are both redrawn (`candidate_seeds.py`), everything else is the frozen script. Points of best-of-n recovery against the pooled coefficient; frozen: the seed of the paper; others: the range over the other seeds; s.d.: the standard deviation over all the seeds, which is the Monte Carlo error of one run's estimate, splits and subsamples together. The intervals of the paper include the split part of it, a bootstrap replicate averaging as many splits as the estimate.

	eight labelled units			half		
	frozen	others	s.d.	frozen	others	s.d.
Exploratory, 19 configurations; few = 4; 4 other seeds						
$D_{\mathrm{grp}}(\mathrm{all})$	+1.2	+0.8 to +1.1	0.2	+1.6	+1.4 to +1.6	0.1
$D_{\mathrm{grp}}(\mathrm{few})$	−1.0	−1.4 to −1.3	0.2	−0.3	−0.2 to +0.3	0.2
$D_W(\mathrm{all})$	+1.6	+1.4 to +1.7	0.1	+1.2	+1.1 to +1.3	0.1
$D_W(\mathrm{few})$	+0.5	−0.0 to +0.4	0.2	+0.6	+0.7 to +0.9	0.1
grp's fall, all to few	+2.1	+2.1 to +2.4	0.1	+1.9	+1.3 to +1.8	0.3
$D_W(\mathrm{few}) - D_{\mathrm{grp}}(\mathrm{few})$	+1.5	+1.3 to +1.7	0.1	+0.9	+0.5 to +1.0	0.2
grp's fall less r_W's	+1.0	+0.6 to +1.0	0.2	+1.3	+0.8 to +1.5	0.2
WMT23 MQM, 3; few = 4; 4 other seeds						
$D_{\mathrm{grp}}(\mathrm{all})$	−0.0	−0.2 to +0.4	0.2	−0.6	−0.4 to −0.2	0.2
$D_{\mathrm{grp}}(\mathrm{few})$	−1.8	−3.0 to −1.7	0.5	−0.1	−0.3 to +0.3	0.2
$D_W(\mathrm{all})$	−0.1	+0.3 to +0.7	0.3	+0.7	+0.7 to +1.0	0.1
$D_W(\mathrm{few})$	−1.1	−1.7 to −0.6	0.4	+1.0	+0.5 to +0.8	0.2
grp's fall, all to few	+1.8	+1.6 to +2.8	0.5	−0.5	−0.6 to −0.2	0.2
$D_W(\mathrm{few}) - D_{\mathrm{grp}}(\mathrm{few})$	+0.7	+0.7 to +1.4	0.4	+1.0	+0.4 to +1.0	0.3
grp's fall less r_W's	+0.8	+0.2 to +0.8	0.2	−0.3	−0.7 to −0.1	0.2
RewardBench 2, 5; few = 2; 4 other seeds						
$D_{\mathrm{grp}}(\mathrm{all})$	−2.4	−3.5 to −2.2	0.6	+1.8	+1.8 to +1.9	0.1
$D_{\mathrm{grp}}(\mathrm{few})$	−5.1	−6.0 to −4.9	0.5	−0.2	−0.4 to +0.4	0.3
$D_W(\mathrm{all})$	+0.9	+0.7 to +1.7	0.4	+1.3	+1.2 to +1.4	0.1
$D_W(\mathrm{few})$	−2.1	−3.1 to −1.8	0.5	+3.4	+2.5 to +3.3	0.4
grp's fall, all to few	+2.8	+1.3 to +3.8	0.9	+2.0	+1.5 to +2.2	0.3
$D_W(\mathrm{few}) - D_{\mathrm{grp}}(\mathrm{few})$	+3.0	+2.2 to +3.4	0.5	+3.5	+2.3 to +3.2	0.4
grp's fall less r_W's	−0.3	−2.2 to −0.2	0.8	+4.1	+2.9 to +3.7	0.4
PRM800K, 2; few = 3; 12 other seeds						
$D_{\mathrm{grp}}(\mathrm{all})$	−7.3	−6.1 to −4.3	0.8	−0.2	−0.6 to −0.1	0.1
$D_{\mathrm{grp}}(\mathrm{few})$	−7.8	−7.5 to −6.0	0.5	−3.3	−3.5 to −1.5	0.6
$D_W(\mathrm{all})$	+0.0	+0.0 to +2.3	0.6	+0.0	+0.0 to +0.0	0.0
$D_W(\mathrm{few})$	−0.8	−1.4 to +0.8	0.7	+0.0	−0.1 to +0.0	0.0
grp's fall, all to few	+0.5	+0.6 to +2.2	0.6	+3.1	+1.4 to +3.3	0.6
$D_W(\mathrm{few}) - D_{\mathrm{grp}}(\mathrm{few})$	+7.0	+5.6 to +7.6	0.6	+3.3	+1.5 to +3.5	0.6
grp's fall less r_W's	−0.3	−0.7 to +1.7	0.8	+3.1	+1.4 to +3.3	0.6
PPE, 5; few = 4; 4 other seeds						
$D_{\mathrm{grp}}(\mathrm{all})$	+1.5	+1.3 to +1.6	0.1	+0.1	+0.0 to +0.1	0.0
$D_{\mathrm{grp}}(\mathrm{few})$	+0.9	+0.7 to +1.3	0.2	−0.1	+0.2 to +0.5	0.2
$D_W(\mathrm{all})$	+1.6	+1.5 to +1.7	0.1	+0.9	+0.9 to +1.0	0.0
$D_W(\mathrm{few})$	+1.1	+0.8 to +1.2	0.2	+0.6	+0.6 to +1.0	0.2
grp's fall, all to few	+0.6	+0.2 to +0.7	0.2	+0.2	−0.3 to −0.1	0.2
$D_W(\mathrm{few}) - D_{\mathrm{grp}}(\mathrm{few})$	+0.2	−0.0 to +0.2	0.1	+0.7	+0.4 to +0.6	0.1
grp's fall less r_W's	+0.0	−0.1 to +0.0	0.1	−0.1	−0.4 to −0.2	0.1

Ordering against choice, by configuration. `scripts/decision/top_pairs.py` bins, on the main run's own draws, the evaluator pairs decisive on the evaluation half by the rank there of the pair's better member and records each rule's selection error in every bin (Table 36, every configuration; Table 35, the sets). The pairs that involve the evaluation half's best evaluator are the $K-1$ pairs of that evaluator, $2/K$ of all pairs and 14% of the decisive pairs over the primary configurations, so a share of the ordering advantage that falls on them says little without that base rate; the table therefore reports the advantage per pair, +3.0 points on those pairs against +5.4 on the rest at half the clusters under best of n (+7.6 against +8.1 under the pairwise choice), +19.5 against +12.0 on WMT23, +0.5 against +1.8 on RewardBench 2, and −13.1 against −2.5 on PRM800K at eight problems, where the five models' ten pairs hold four with the best. A rule selects the best evaluator when every one of those $K-1$ pairs is ordered correctly on the same draw, so the selection rate is a joint event and not a function of the mean error on them, which is that evaluator's mean rank less one over $K-1$: the two move the other way on 2 of the primary configurations and on 3 of the 27 primary, WMT23 and RewardBench 2 configurations, WMT23 en-de among them, where the advantage on those pairs is +3.1 points and the grouped mean selects the best metric on 3% of draws against 20%. What a miss costs is the regret of the evaluator selected instead, which depends on the ordering among the other evaluators. The utility difference of Table 5 is therefore split, exactly and per configuration,

Table 27: A fixed number of labels $N = b \times m$ split between labelled clusters b and candidates kept in each m, on PPE (five sets, best-of-16 accuracy, %), read after the fact and not among the protocol's hypotheses; Tables 28 and 29 give the exploratory configurations and WMT23. Per cell: the utility of the evaluator each rule selects and the paired difference against pooling on the same subsampled clusters; in small type, the paired difference between the cell with the most candidates a cluster and each other cell of the same N, for the utility of each rule and for the two differences against pooling, the last of which is the contrast in the quantity the protocol grades. Intervals: the estimate ± 1.96 standard deviations of the frozen script's bootstrap rerun at these budgets (`candidate_alloc.py`; the same resampled corpora and splits in every cell, a smaller budget a nested subset of a larger one's clusters). Budgets above 24 prompts are left out: the selection half of a resampled PPE pool holds about thirty distinct prompts, so few of its splits reach 32. Estimates average 500 selection draws and 20 subsamples of responses; a bootstrap replicate averages 300 draws, and the intervals are conditional on the subsamples.

N	b	m	r_T	r_W	grp	$r_W - r_T$	grp $- r_T$
24	4	6	53.7	54.7	54.4	+1.0 [+0.4, +1.5]	+0.6 [−0.0, +1.3]
	6	4	53.6	54.6	54.6	+1.0 [+0.4, +1.6]	+1.0 [+0.3, +1.7]
	8	3	53.8	54.5	54.3	+0.8 [+0.0, +1.5]	+0.5 [−0.3, +1.3]
	12	2	53.6	54.1	53.6	+0.5 [−0.4, +1.4]	+0.0 [−0.9, +0.9]
	4×6 − 6×4		+0.1 [−0.3, +0.5]	+0.1 [−0.4, +0.5]	−0.2 [−0.7, +0.2]	−0.0 [−0.6, +0.5]	−0.3 [−1.0, +0.3]
	4×6 − 8×3		−0.0 [−0.5, +0.5]	+0.2 [−0.4, +0.7]	+0.1 [−0.5, +0.6]	+0.2 [−0.4, +0.8]	+0.1 [−0.5, +0.7]
	4×6 − 12×2		+0.2 [−0.4, +0.8]	+0.6 [+0.1, +1.1]	+0.8 [+0.2, +1.4]	+0.4 [−0.3, +1.2]	+0.6 [−0.2, +1.4]
32	4	8	53.8	55.0	54.8	+1.1 [+0.6, +1.7]	+1.0 [+0.3, +1.6]
	8	4	53.9	55.0	54.8	+1.1 [+0.3, +1.8]	+0.9 [+0.1, +1.7]
	16	2	54.1	54.7	54.2	+0.6 [−0.4, +1.5]	+0.1 [−0.9, +1.1]
	4×8 − 8×4		−0.1 [−0.6, +0.4]	−0.0 [−0.5, +0.4]	−0.0 [−0.5, +0.5]	+0.1 [−0.5, +0.6]	+0.1 [−0.5, +0.7]
	4×8 − 16×2		−0.2 [−1.0, +0.5]	+0.3 [−0.2, +0.8]	+0.6 [+0.1, +1.2]	+0.6 [−0.2, +1.3]	+0.9 [+0.0, +1.7]
48	6	8	54.3	55.7	55.5	+1.4 [+0.6, +2.1]	+1.2 [+0.3, +2.0]
	8	6	54.5	55.8	55.5	+1.3 [+0.5, +2.1]	+1.0 [+0.2, +1.9]
	12	4	54.8	55.8	55.6	+1.1 [+0.2, +1.9]	+0.8 [−0.1, +1.7]
	16	3	55.0	55.9	55.4	+0.9 [−0.1, +2.0]	+0.4 [−0.6, +1.5]
	24	2	54.6	55.3	55.0	+0.7 [−0.4, +1.8]	+0.4 [−0.7, +1.6]
	6×8 − 8×6		−0.2 [−0.6, +0.2]	−0.1 [−0.5, +0.2]	−0.1 [−0.5, +0.3]	+0.1 [−0.4, +0.6]	+0.1 [−0.4, +0.6]
	6×8 − 12×4		−0.5 [−1.0, +0.1]	−0.1 [−0.6, +0.3]	−0.1 [−0.6, +0.3]	+0.3 [−0.3, +0.9]	+0.4 [−0.3, +1.0]
	6×8 − 16×3		−0.7 [−1.3, −0.1]	−0.2 [−0.7, +0.3]	+0.1 [−0.4, +0.5]	+0.5 [−0.3, +1.2]	+0.8 [+0.0, +1.5]
	6×8 − 24×2		−0.3 [−1.3, +0.7]	+0.4 [−0.4, +1.1]	+0.4 [−0.3, +1.2]	+0.7 [−0.2, +1.6]	+0.7 [−0.2, +1.7]
64	4	16	54.3	55.5	55.5	+1.2 [+0.5, +1.9]	+1.2 [+0.5, +2.0]
	8	8	54.7	56.1	56.0	+1.4 [+0.6, +2.2]	+1.3 [+0.4, +2.2]
	16	4	55.1	56.3	55.9	+1.2 [+0.2, +2.2]	+0.8 [−0.3, +1.9]
	4×16 − 8×8		−0.4 [−0.9, +0.1]	−0.6 [−1.1, −0.2]	−0.5 [−0.8, −0.1]	−0.2 [−0.8, +0.3]	−0.0 [−0.6, +0.5]
	4×16 − 16×4		−0.8 [−1.6, −0.1]	−0.8 [−1.4, −0.2]	−0.4 [−1.0, +0.2]	+0.0 [−0.8, +0.9]	+0.4 [−0.4, +1.3]
96	6	16	54.7	56.3	56.1	+1.6 [+0.8, +2.4]	+1.4 [+0.5, +2.3]
	12	8	55.2	56.7	56.3	+1.5 [+0.5, +2.5]	+1.1 [+0.0, +2.2]
	16	6	55.5	56.8	56.4	+1.3 [+0.2, +2.3]	+0.9 [−0.3, +2.0]
	24	4	55.7	56.8	56.3	+1.1 [−0.2, +2.3]	+0.6 [−0.6, +1.9]
	6×16 − 12×8		−0.5 [−1.1, +0.0]	−0.4 [−0.8, +0.0]	−0.2 [−0.6, +0.2]	+0.1 [−0.5, +0.8]	+0.3 [−0.3, +1.0]
	6×16 − 16×6		−0.8 [−1.5, −0.2]	−0.5 [−0.9, −0.0]	−0.3 [−0.7, +0.2]	+0.4 [−0.4, +1.1]	+0.6 [−0.2, +1.3]
	6×16 − 24×4		−1.0 [−2.1, +0.1]	−0.4 [−1.3, +0.4]	−0.2 [−1.0, +0.6]	+0.6 [−0.4, +1.5]	+0.8 [−0.2, +1.8]
128	8	16	55.1	56.7	56.6	+1.6 [+0.7, +2.5]	+1.5 [+0.5, +2.5]
	16	8	55.8	57.2	56.7	+1.4 [+0.2, +2.6]	+0.9 [−0.3, +2.1]
	8×16 − 16×8		−0.6 [−1.2, −0.1]	−0.5 [−0.8, −0.1]	−0.1 [−0.5, +0.3]	+0.2 [−0.4, +0.8]	+0.6 [−0.1, +1.2]

Table 28: As Table 27, on the 19 exploratory configurations (best-of-n recovery, %). The selection half of a resampled HANNA, SummEval or RealSumm corpus, 11 of the 19 configurations, holds about thirty distinct clusters, so few of its splits reach 32 and 24 clusters are most of what it holds; the cells with 32 clusters are printed as estimates on the corpus itself, without intervals. FRANK's clusters hold five candidates at the median, so its cells with more candidates hold fewer labels than N. Estimates average 300 splits and 20 subsamples of candidates; intervals are over 200 corpus-bootstrap replicates of 300 splits each, conditional on those subsamples, and were not run through the calibration.

N	b	m	r_T	r_W	grp	$r_W - r_T$	grp $- r_T$
24	4	6	40.3	40.4	39.6	+0.1 [−0.6, +0.8]	−0.8 [−1.6, +0.1]
	6	4	40.2	40.2	38.8	+0.1 [−0.7, +0.8]	−1.3 [−2.2, −0.5]
	8	3	40.6	40.3	38.3	−0.3 [−1.1, +0.5]	−2.2 [−3.2, −1.3]
	12	2	40.4	39.2	36.5	−1.1 [−1.9, −0.3]	−3.8 [−5.0, −2.7]
	4×6 − 6×4		+0.2 [−0.3, +0.6]	+0.2 [−0.4, +0.7]	+0.7 [+0.1, +1.3]	+0.0 [−0.5, +0.6]	+0.6 [−0.1, +1.3]
	4×6 − 8×3		−0.3 [−0.8, +0.3]	+0.1 [−0.5, +0.7]	+1.2 [+0.6, +1.9]	+0.4 [−0.3, +1.1]	+1.5 [+0.7, +2.3]
	4×6 − 12×2		−0.0 [−0.6, +0.5]	+1.2 [+0.6, +1.8]	+3.0 [+2.2, +3.9]	+1.2 [+0.5, +1.9]	+3.1 [+2.1, +4.1]
32	4	8	41.1	41.5	40.8	+0.4 [−0.3, +1.1]	−0.3 [−1.1, +0.4]
	8	4	40.8	41.3	39.9	+0.5 [−0.3, +1.3]	−1.0 [−1.9, −0.0]
	16	2	41.4	40.1	37.7	−1.4 [−2.2, −0.5]	−3.8 [−5.0, −2.5]
	4×8 − 8×4		+0.3 [−0.2, +0.8]	+0.2 [−0.3, +0.7]	+1.0 [+0.4, +1.5]	−0.1 [−0.7, +0.4]	+0.7 [−0.0, +1.4]
	4×8 − 16×2		−0.3 [−0.9, +0.3]	+1.4 [+0.8, +2.1]	+3.1 [+2.3, +4.0]	+1.7 [+1.0, +2.5]	+3.4 [+2.4, +4.5]
48	6	8	42.3	43.1	42.3	+0.9 [+0.0, +1.7]	+0.0 [−0.9, +0.9]
	8	6	42.2	43.1	42.1	+0.9 [+0.1, +1.8]	−0.1 [−1.0, +0.8]
	12	4	42.2	42.9	41.6	+0.7 [−0.2, +1.6]	−0.6 [−1.5, +0.3]
	16	3	42.7	42.9	41.1	+0.1 [−0.8, +1.0]	−1.7 [−2.7, −0.6]
	24	2	42.5	41.8	39.3	−0.7 [−1.7, +0.3]	−3.1 [−4.5, −1.8]
	6×8 − 8×6		+0.1 [−0.3, +0.4]	−0.0 [−0.5, +0.4]	+0.1 [−0.4, +0.6]	−0.1 [−0.6, +0.4]	+0.1 [−0.5, +0.6]
	6×8 − 12×4		+0.0 [−0.4, +0.5]	+0.2 [−0.3, +0.6]	+0.6 [+0.1, +1.2]	+0.1 [−0.4, +0.7]	+0.6 [−0.0, +1.2]
	6×8 − 16×3		−0.5 [−1.1, +0.1]	+0.2 [−0.3, +0.8]	+1.2 [+0.6, +1.8]	+0.7 [+0.1, +1.3]	+1.7 [+0.8, +2.5]
	6×8 − 24×2		−0.2 [−0.9, +0.5]	+1.3 [+0.7, +2.0]	+2.9 [+2.1, +3.8]	+1.5 [+0.8, +2.3]	+3.1 [+2.1, +4.2]
64	8	8	43.0	44.1	43.0	+1.0 [+0.1, +2.0]	+0.0 [−0.9, +1.0]
	16	4	43.0	43.8	42.7	+0.8 [−0.1, +1.7]	−0.3 [−1.3, +0.6]
	32	2	43.3	42.9	40.5	−0.5	−2.8
	8×8 − 16×4		−0.0 [−0.5, +0.5]	+0.2 [−0.2, +0.7]	+0.3 [−0.2, +0.9]	+0.2 [−0.3, +0.8]	+0.3 [−0.3, +1.0]
	8×8 − 32×2		−0.3	+1.2	+2.5	+1.5	+2.8
96	12	8	43.7	44.9	44.2	+1.2 [+0.1, +2.3]	+0.4 [−0.6, +1.5]
	16	6	43.9	45.1	44.2	+1.2 [+0.1, +2.2]	+0.2 [−0.9, +1.3]
	24	4	44.1	45.0	43.8	+0.9 [−0.2, +2.0]	−0.3 [−1.4, +0.8]
	32	3	44.2	44.9	43.4	+0.7	−0.8
	12×8 − 16×6		−0.2 [−0.6, +0.2]	−0.2 [−0.6, +0.2]	+0.0 [−0.4, +0.4]	+0.0 [−0.4, +0.4]	+0.2 [−0.3, +0.7]
	12×8 − 24×4		−0.4 [−0.8, +0.1]	−0.1 [−0.6, +0.3]	+0.3 [−0.2, +0.9]	+0.3 [−0.3, +0.8]	+0.7 [+0.0, +1.4]
	12×8 − 32×3		−0.5	+0.0	+0.7	+0.5	+1.2
128	16	8	44.4	45.6	45.0	+1.2 [+0.0, +2.4]	+0.6 [−0.6, +1.8]
	32	4	44.8	45.7	44.7	+0.9	−0.2
	16×8 − 32×4		−0.4	−0.1	+0.3	+0.4	+0.7

Table 29: As Table 27, on the three WMT23 MQM configurations (best-of-n recovery, %), whose selection halves hold 198 to 535 segments, so every budget of the grid is reached in every replicate. Estimates average 300 splits and 20 subsamples of candidates; intervals are over 200 corpus-bootstrap replicates of 300 splits each, conditional on those subsamples, and were not run through the calibration.

N	b	m	r_T	r_W	grp	$r_W - r_T$	grp $- r_T$
24	4	6	47.2	46.3	45.6	−0.9 [−1.9, +0.1]	−1.7 [−3.0, −0.3]
	6	4	47.9	46.2	44.1	−1.7 [−2.8, −0.6]	−3.8 [−5.3, −2.3]
	8	3	48.0	45.6	44.3	−2.3 [−3.6, −1.1]	−3.7 [−5.2, −2.3]
	12	2	48.3	46.3	43.2	−2.0 [−3.4, −0.5]	−5.2 [−6.8, −3.5]
	4×6 − 6×4		−0.7 [−1.7, +0.4]	+0.1 [−1.0, +1.3]	+1.5 [+0.2, +2.8]	+0.8 [−0.5, +2.1]	+2.1 [+0.6, +3.7]
	4×6 − 8×3		−0.7 [−1.9, +0.5]	+0.7 [−0.5, +1.9]	+1.3 [+0.0, +2.6]	+1.4 [−0.0, +2.9]	+2.0 [+0.4, +3.7]
	4×6 − 12×2		−1.1 [−2.3, +0.1]	−0.0 [−1.4, +1.4]	+2.4 [+1.0, +3.9]	+1.1 [−0.6, +2.7]	+3.5 [+1.6, +5.4]
32	4	8	47.8	47.0	46.5	−0.8 [−1.9, +0.3]	−1.3 [−2.7, +0.1]
	8	4	48.1	47.1	46.3	−1.1 [−2.2, +0.1]	−1.8 [−3.2, −0.4]
	16	2	49.1	46.8	44.3	−2.3 [−3.7, −0.9]	−4.8 [−6.4, −3.1]
	4×8 − 8×4		−0.3 [−1.4, +0.8]	−0.0 [−1.2, +1.1]	+0.2 [−1.1, +1.5]	+0.3 [−1.1, +1.7]	+0.5 [−1.1, +2.1]
	4×8 − 16×2		−1.2 [−2.5, +0.0]	+0.2 [−1.1, +1.6]	+2.2 [+0.8, +3.6]	+1.5 [−0.1, +3.0]	+3.5 [+1.6, +5.3]
48	6	8	48.8	48.9	47.8	+0.1 [−0.9, +1.1]	−1.0 [−2.3, +0.3]
	8	6	49.4	48.7	48.3	−0.6 [−1.6, +0.4]	−1.0 [−2.3, +0.3]
	12	4	49.1	48.9	47.8	−0.2 [−1.4, +0.9]	−1.3 [−2.8, +0.3]
	16	3	49.4	48.2	47.3	−1.1 [−2.4, +0.1]	−2.1 [−3.8, −0.5]
	24	2	50.1	48.3	45.9	−1.8 [−3.3, −0.3]	−4.2 [−5.9, −2.5]
	6×8 − 8×6		−0.5 [−1.4, +0.3]	+0.2 [−0.7, +1.1]	−0.5 [−1.5, +0.5]	+0.7 [−0.5, +1.9]	+0.0 [−1.3, +1.4]
	6×8 − 12×4		−0.3 [−1.3, +0.7]	+0.1 [−1.0, +1.1]	−0.0 [−1.2, +1.1]	+0.3 [−1.0, +1.7]	+0.3 [−1.3, +1.8]
	6×8 − 16×3		−0.6 [−1.7, +0.6]	+0.7 [−0.5, +1.8]	+0.6 [−0.5, +1.7]	+1.3 [−0.2, +2.7]	+1.1 [−0.5, +2.8]
	6×8 − 24×2		−1.3 [−2.5, −0.1]	+0.6 [−0.5, +1.8]	+1.9 [+0.7, +3.1]	+1.9 [+0.3, +3.5]	+3.2 [+1.5, +4.9]
64	8	8	49.8	50.1	49.1	+0.3 [−0.6, +1.2]	−0.6 [−2.0, +0.7]
	16	4	49.8	49.6	48.4	−0.2 [−1.4, +1.1]	−1.3 [−2.9, +0.3]
	32	2	50.6	49.6	46.7	−1.0 [−2.5, +0.5]	−3.8 [−5.7, −2.0]
	8×8 − 16×4		−0.0 [−1.0, +1.0]	+0.5 [−0.4, +1.3]	+0.7 [−0.3, +1.6]	+0.5 [−0.7, +1.6]	+0.7 [−0.7, +2.0]
	8×8 − 32×2		−0.8 [−2.0, +0.4]	+0.5 [−0.6, +1.5]	+2.4 [+1.1, +3.6]	+1.3 [−0.3, +2.8]	+3.2 [+1.5, +4.9]
96	12	8	50.7	51.6	50.5	+0.9 [−0.1, +1.8]	−0.2 [−1.6, +1.2]
	16	6	50.0	51.1	50.6	+1.1 [+0.1, +2.2]	+0.7 [−0.8, +2.2]
	24	4	51.1	51.4	49.5	+0.2 [−1.0, +1.4]	−1.7 [−3.4, +0.1]
	32	3	52.0	50.7	49.7	−1.4 [−2.7, −0.0]	−2.3 [−4.0, −0.7]
	12×8 − 16×6		+0.8 [+0.0, +1.5]	+0.5 [−0.2, +1.2]	−0.1 [−1.0, +0.7]	−0.3 [−1.2, +0.7]	−0.9 [−2.0, +0.2]
	12×8 − 24×4		−0.4 [−1.4, +0.5]	+0.2 [−0.6, +1.0]	+1.0 [+0.1, +1.9]	+0.6 [−0.5, +1.7]	+1.4 [+0.1, +2.8]
	12×8 − 32×3		−1.3 [−2.2, −0.4]	+0.9 [+0.0, +1.8]	+0.8 [−0.2, +1.8]	+2.2 [+1.1, +3.4]	+2.1 [+0.8, +3.5]
128	16	8	51.3	52.5	51.6	+1.2 [+0.2, +2.2]	+0.4 [−1.1, +1.8]
	32	4	52.2	51.7	50.9	−0.5 [−1.7, +0.7]	−1.3 [−3.1, +0.5]
	16×8 − 32×4		−0.9 [−1.7, −0.1]	+0.7 [−0.1, +1.5]	+0.7 [−0.1, +1.5]	+1.6 [+0.6, +2.7]	+1.6 [+0.6, +2.7]

Table 30: The addendum: candidates removed from the two panels whose clusters hold four and six, with its four graded contrasts (`prereg/PROTOCOL_candidates_addendum.md`, frozen and time-stamped before it ran). Paired differences against the pooled coefficient computed on the same subsampled clusters, in points of best-of-n recovery on the evaluation half, the estimate ± 1.96 corpus-bootstrap standard deviations; a blank cell is a number of candidates the panel does not hold or the addendum did not run. With two candidates the grouped mean, the per-cluster Pearson mean and the pairwise accuracy are one statistic. The grades are descriptive, without correction and outside the calibration; the leave-one-subset-out of A1 is in the text. Estimates average 300 splits and 20 subsamples of candidates; intervals are over 200 corpus-bootstrap replicates of 300 splits each, conditional on those subsamples, and were not run through the calibration.

candidates kept, m	2	3	4	all
RewardBench 2, five subsets, four completions a prompt, eight labelled clusters				
clusters that ask a question	3.9	6.0		8.0
$r_W - r_T$	−2.1 [−3.2, −1.0]	+0.0 [−1.0, +1.1]		+0.9 [−0.2, +2.1]
grp $- r_T$	−5.1 [−6.5, −3.8]	−5.0 [−6.2, −3.7]		−2.4 [−3.8, −1.0]
per-cluster Pearson mean $- r_T$	−5.1 [−6.5, −3.8]	−1.6 [−2.7, −0.4]		+0.6 [−0.5, +1.7]
pairwise accuracy $- r_T$	−5.1 [−6.5, −3.8]	−5.0 [−6.2, −3.8]		−3.0 [−4.4, −1.6]
best of $n - r_T$	−5.1 [−6.5, −3.8]	−4.8 [−6.0, −3.6]		−3.4 [−4.9, −2.0]
RewardBench 2, five subsets, four completions a prompt, half the clusters				
clusters that ask a question	87.4	132.1		176.0
$r_W - r_T$	+3.4 [+1.6, +5.1]	+1.1 [−0.5, +2.7]		+1.3 [−0.1, +2.6]
grp $- r_T$	−0.2 [−2.1, +1.8]	+1.1 [−0.8, +3.1]		+1.8 [+0.2, +3.5]
per-cluster Pearson mean $- r_T$	−0.2 [−2.1, +1.8]	+2.0 [+0.0, +3.9]		+2.1 [+0.5, +3.7]
pairwise accuracy $- r_T$	−0.2 [−2.1, +1.8]	+1.1 [−0.8, +3.1]		+1.8 [+0.2, +3.4]
best of $n - r_T$	−0.2 [−2.1, +1.8]	+0.6 [−1.4, +2.6]		+1.2 [−0.6, +3.0]
A1, D_{grp}(all) - $D_{\mathrm{grp}}(2) > 0$: +2.8 [+1.2, +4.4], supported				
A2, $D_W(2)$ - $D_{\mathrm{grp}}(2) > 0$: +3.0 [+1.8, +4.2], supported				
PRM800K, two labels, six solutions a problem, eight labelled clusters				
clusters that ask a question	4.9	5.7	6.1	6.5
$r_W - r_T$	−2.6 [−4.2, −1.0]	−0.8 [−2.4, +0.8]	−0.6 [−2.0, +0.8]	+0.0 [−1.4, +1.5]
grp $- r_T$	−7.5 [−9.9, −5.1]	−7.8 [−10.1, −5.5]	−7.6 [−9.7, −5.6]	−7.3 [−9.6, −5.0]
per-cluster Pearson mean $- r_T$	−7.5 [−9.9, −5.1]	−6.5 [−8.7, −4.3]	−5.4 [−7.2, −3.6]	−3.7 [−5.6, −1.7]
pairwise accuracy $- r_T$	−7.5 [−9.9, −5.1]	−8.2 [−10.5, −5.9]	−7.5 [−9.5, −5.5]	−6.7 [−8.9, −4.5]
best of $n - r_T$	−8.1 [−10.4, −5.8]	−7.9 [−10.2, −5.6]	−6.5 [−8.5, −4.5]	−4.8 [−7.1, −2.6]
PRM800K, two labels, six solutions a problem, half the clusters				
clusters that ask a question	135.4	160.7	173.1	183.1
$r_W - r_T$	−0.1 [−1.2, +1.0]	+0.0 [−0.6, +0.6]	+0.0 [−0.3, +0.3]	+0.0 [−0.1, +0.1]
grp $- r_T$	−3.5 [−6.3, −0.7]	−3.3 [−6.2, −0.4]	−1.3 [−4.1, +1.5]	−0.2 [−2.9, +2.5]
per-cluster Pearson mean $- r_T$	−3.5 [−6.3, −0.7]	−2.8 [−5.5, −0.2]	−0.8 [−3.0, +1.4]	−0.1 [−1.7, +1.6]
pairwise accuracy $- r_T$	−3.5 [−6.3, −0.7]	−3.5 [−6.5, −0.6]	−1.9 [−5.0, +1.1]	−0.3 [−3.0, +2.4]
best of $n - r_T$	−2.5 [−5.1, +0.1]	−0.5 [−2.0, +1.1]	+0.0 [−0.8, +0.8]	+0.0 [−0.3, +0.3]
A3, D_{grp}(all) - $D_{\mathrm{grp}}(3) > 0$: +0.5 [−1.6, +2.6], consistent				
A4, $D_W(3)$ - $D_{\mathrm{grp}}(3) > 0$: +7.0 [+5.0, +9.0], supported				

Table 31: The grouped mean against pooling, read after the fact: the correlation r across configurations of its ordering advantage (points of selection error over evaluator pairs) with the utility gain of the evaluator it selects (points), and family means. Gains at eight clusters and at half carry the corpus-bootstrap interval of the family mean; half is 48 to 249 clusters outside machine translation and 230 to 939 segments on it. The last two columns are the three MQM language pairs of WMT24, the primary set of the third protocol, which no correlation of the table includes (half is 241 to 306 segments; grades in Table 48).

		r of advantage and gain		WMT22 and WMT23 (6)		the other 21		WMT24 (3), frozen test	
payoff	clusters	all 27	without MT	advantage	gain	advantage	gain	advantage	gain
best of n	8	0.68	0.73	+2.3	+1.2 [+0.3, +2.1]	+0.7	+0.1 [−0.8, +1.1]	+3.4	+1.4
	32	0.48	0.60	+8.7	+2.0	+1.7	+1.3	+6.9	+0.7
	half	0.19	0.51	+14.1	−0.1 [−1.9, +1.7]	+2.7	+1.8 [+0.1, +3.6]	+10.8	−2.8 [−9.2, +3.7]
pairwise choice	8	0.72	0.75	−0.3	+0.1 [−0.1, +0.3]	+1.3	+0.2 [−0.1, +0.6]	+4.0	+0.6
	32	0.29	0.36	+4.9	+0.3	+4.3	+0.9	+9.2	+0.6
	half	0.04	0.24	+13.3	−0.1 [−0.5, +0.3]	+5.8	+1.1 [+0.6, +1.5]	+15.9	−0.2 [−1.1, +0.7]

as $(h_G - h_T)\,\bar{L} + (1 - \bar{h})\,(L_T - L_G)$, with h_r the rate at which rule r selects the best evaluator, L_r its mean regret when it does not, and $\bar{h}$, $\bar{L}$ the means of the two: the naming part and the loss part of Tables 35 and 36, whose set rows are means over configurations of the per-configuration parts. That split is the symmetric member of a family: $U_G - U_T = (1 - h_T)L_T - (1 - h_G)L_G$ also equals $(h_G - h_T)L_T + (1 - h_G)(L_T - L_G)$ and $(h_G - h_T)L_G + (1 - h_T)(L_T - L_G)$, under

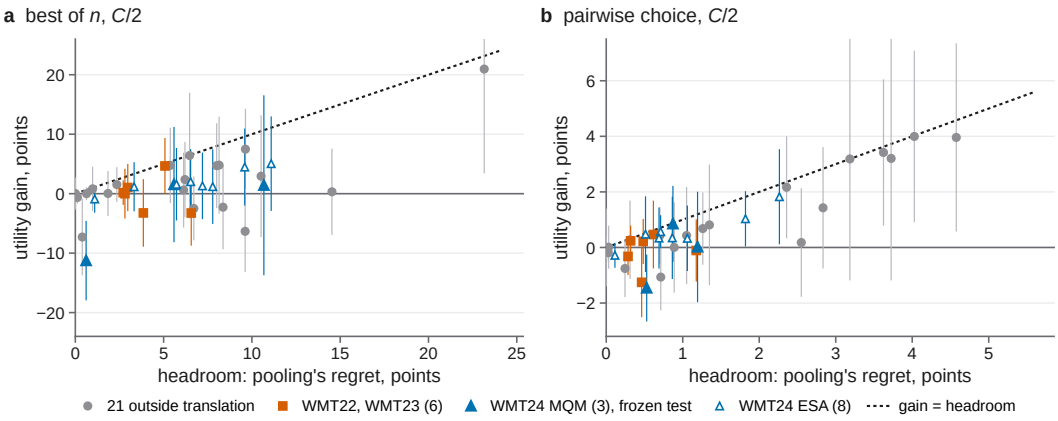


Figure 9: **Gain against headroom.** The utility gain at half the clusters against pooling's regret on the evaluation half (the headroom), under best of n and under the pairwise choice, marks as in Figure 8. The gain is that regret less the grouped mean's own, so no point lies above the diagonal and the correlation across configurations is set by the two regrets' spreads and their correlation (Section 7). Read after the fact (`scripts/decision/r16_headroom.py`).

Table 32: Where the grouped mean's gain over pooling comes from, by configuration, under best of n at half the clusters against the out-of-sample target (the complement; PPE: the fixed evaluation set), 300 splits (PPE 500 draws). Read off the corpus once: whether the evaluator the pooled coefficient names from every cluster is the payoff's best there (lim.) and the margin between the best and the second-best evaluator as a share of the span between the best and the mean one. Then the share of draws on which the two rules select different evaluators (dis.), the payoff difference of their selected evaluators in points, grouped minus pooled, which is dis. times the difference given disagreement, and the share of disagreeing draws the grouped mean's evaluator wins; for each rule the share of draws naming the target's best evaluator, the regret share given a miss, and the regret share. Percent unless marked.

		Corpus			grp against r_T			Selects best		Loss if missed		Regret share	
Configuration	C	Ev.	lim.	margin	dis.	diff., pts	wins	r_T	grp	r_T	grp	r_T	grp
speechocean762 accuracy	125	9	✓	42	0	+0.0	–	83	83	24	24	4	4
speechocean762 fluency	125	9	–	16	73	+2.9	64	9	20	102	83	93	66
speechocean762 prosody	125	9	–	62	55	+4.7	80	29	57	78	59	55	25
speechocean762 total	125	9	✓	9	1	+0.0	67	62	62	27	26	10	10
FRANK factuality	499	11	✓	82	33	+6.4	100	65	97	117	22	41	1
HANNA coherence	96	19	–	3	23	+0.6	65	20	26	23	22	18	16
HANNA complexity	96	19	–	5	94	+4.8	78	10	36	25	15	23	9
HANNA empathy	96	19	–	22	36	−2.5	31	17	13	28	36	23	31
HANNA engagement	96	19	✓	2	76	−2.3	35	22	7	33	35	26	33
HANNA relevance	96	19	–	48	94	+21.0	100	0	92	63	69	63	6
HANNA surprise	96	19	–	23	80	+0.3	56	2	2	59	58	58	56
RealSumm LitePyramid	100	16	–	25	74	+2.4	71	11	30	59	46	53	32
SummEval coherence	100	16	✓	31	54	−7.3	1	92	41	18	46	1	27
SummEval consistency	100	16	–	8	99	+4.8	92	10	76	16	7	14	2
SummEval fluency	100	16	–	8	44	−6.3	0	30	0	36	42	25	42
SummEval relevance	100	16	✓	48	0	+0.0	–	100	100	–	–	0	0
WMT22 en-de	1315	18	–	8	65	−0.0	54	31	41	25	29	17	17
WMT22 en-ru	1315	17	✓	10	47	+1.0	66	26	34	27	20	20	13
WMT22 zh-en	1875	18	–	13	41	+0.2	57	17	21	28	28	24	22
PRM800K, answer label	458	5	✓	85	0	+0.0	–	100	100	–	–	0	0
PRM800K, human step label	457	5	✓	91	1	−0.4	0	100	99	–	111	0	1
PPE MMLU-Pro	96	6	✓	31	67	−3.5	7	89	32	25	64	3	44
PPE MATH	96	6	–	43	25	+0.4	63	79	86	46	49	10	7
PPE GPQA	96	6	✓	20	41	+0.0	67	23	43	26	33	20	19
PPE IFEval	96	6	✓	81	38	+4.3	88	5	29	87	84	83	60
PPE MBPP-Plus	96	6	✓	98	1	+0.6	100	99	100	105	–	1	0

which the primary parts at half the clusters read +1.7 and −0.1, and +1.1 and +0.5 (+0.6 and −0.0, −0.2 and +0.8 without HANNA relevance; RewardBench 2 +1.2 and +0.7, +0.5 and +1.3; WMT23 +1.0 and −1.6, −0.0 and −0.6), a set row's parts moving by up to 0.5 points between members and a configuration's by 2.7, with the larger part changing on 10 of 34 configurations, so the rates and the losses are the quantities, every part is a point estimate without an interval, and a part's size or sign is read only where every member agrees. Over the primary configurations the split is +1.4

Table 33: The direction of the gain and the between-cluster share of the selected evaluators' own scores, best of n at half the clusters, every primary and validation configuration. pooled and grouped: the evaluator each rule selects when it reads every cluster of the corpus (its limit choice); η_p^2: the between-cluster share of the variance of that evaluator's scores under the decision's clustering, the square of the η_p of Equation 1, which needs no labels; $\Delta\eta_p^2$: pooled less grouped; gain: the paired utility difference of the grouped mean's evaluator against pooling's in points of recovery; dis.: the share of draws on which the two select different evaluators; win: the share of those draws the grouped mean's evaluator wins; 300 splits. Over the 13 configurations whose limit choices differ and whose gain has a readable sign at the printed precision, the signs of $\Delta\eta_p^2$ and of the gain are opposite on 11 (two-sided binomial $p = 0.022$, the configurations treated as independent); over the 14 configurations whose limit choices differ, Spearman($\Delta\eta_p^2$, gain) $= -0.44$ ($n = 14$, exact permutation $p = 0.12$; on $\Delta\eta_p$ itself -0.46), and over the 13 with a readable sign -0.38 ($p = 0.19$); an association read after the fact on the configurations that also pay, whose use as a rule is tested in Table 34.

Configuration	pooled selects	grouped selects	η_p^2 pooled	η_p^2 grouped	$\Delta\eta_p^2$	gain	dis. %	win
so762 accuracy	hmamba	same	0.72	0.72	–	+0.0	0	–
so762 fluency	hmamba	cluster_mean_condi...	0.77	0.92	−0.15	+2.9	73	0.64
so762 prosodic	hmamba	cluster_mean_condi...	0.77	0.94	−0.17	+4.7	55	0.80
so762	hmamba	same	0.74	0.74	–	+0.0	1	0.67
FRANK factuality	BertScore F1 Art	same	0.69	0.69	–	+6.4	33	1.00
HANNA coherence	OrcaPlatypus_task1	same	0.21	0.21	–	+0.6	23	0.65
HANNA complexity	Beluga-13B_task1	METEOR	0.13	0.39	−0.26	+4.8	94	0.78
HANNA empathy	Beluga-13B_task1	same	0.18	0.18	–	−2.5	36	0.31
HANNA engagement	OrcaPlatypus_task1	METEOR	0.21	0.39	−0.18	−2.3	76	0.35
HANNA relevance	OrcaPlatypus_task1	BARTScore-SP	0.33	0.80	−0.47	+21.0	94	1.00
HANNA surprise	Beluga-13B_task1	METEOR	0.15	0.39	−0.24	+0.3	80	0.56
RealSumm	bartscore_repo:bart...	bartscore_repo:bart...	0.68	0.71	−0.03	+2.4	74	0.71
SummEval coherence	geval_gpt4	unieval	0.19	0.09	+0.10	−7.3	54	0.01
SummEval consistency	unieval	geval_gpt4	0.08	0.12	−0.05	+4.8	99	0.92
SummEval fluency	geval_gpt4	same	0.14	0.14	–	−6.3	44	0.00
SummEval relevance	geval_gpt4	same	0.19	0.19	–	+0.0	0	–
WMT22 ende	metricx_xxl_MQM_2...	metricx_xl_MQM_20...	0.66	0.64	+0.01	−0.0	65	0.54
WMT22 enru	metricx_xxl_MQM_2...	same	0.59	0.59	–	+1.0	47	0.66
WMT22 zhen	COMET-22-refA	same	0.80	0.80	–	+0.2	41	0.57
WMT23 ende	XCOMET-Ensemble-refA	mbr-metricx-qe-src	0.75	0.72	+0.03	−3.2	93	0.20
WMT23 heen	XCOMET-Ensemble-refB	same	0.71	0.71	–	−3.3	79	0.30
WMT23 zhen	XCOMET-Ensemble-refA	GEMBA-MQM-src	0.73	0.40	+0.33	+4.7	99	0.91
RB2 factuality	Skywork-Reward-V2-Ll...	same	0.46	0.46	–	+0.8	8	0.91
RB2 precise if	Skywork-Reward-V2-Ll...	same	0.79	0.79	–	+0.0	0	–
RB2 math	open_instruct_dev-...	Skywork-Reward-V2-Ll...	0.37	0.41	−0.04	+7.5	78	0.91
RB2 safety	INF-ORM-Llama3.1-70B	RAMO-Llama3.1-8B	0.16	0.21	−0.05	+1.5	89	0.85
RB2 focus	Skywork-Reward-V2-Ll...	same	0.21	0.21	–	−0.6	19	0.00

and +0.2 at half the clusters and +0.5 and +0.6 at eight; on WMT23 +0.5 and −1.1, the loss part carrying the sign on en-de (−2.2) and he-en (−3.2); on RewardBench 2 +0.8 and +1.0 at half and −0.4 and −2.0 at eight; on PRM800K at eight problems −7.2 and −0.1. The ordering advantage over all pairs and the utility difference agree in sign on 13 of the 15 primary configurations at half the clusters and on 13 of 17 at eight; counting as a parting a sign difference or one quantity within a point of zero while the other is beyond two, 8 configurations part at half: HANNA engagement (+0.1, −2.3), HANNA surprise (−0.8, +0.3), SummEval coherence (−0.5, −7.3), SummEval consistency (+0.7, +4.8), SummEval fluency (−0.1, −6.3), SummEval relevance (−6.8, +0.0), wmt22 ende (+14.2, −0.0), wmt22 zhen (+18.7, +0.2) (ordering advantage and utility difference in points). The reference of the split is the evaluation half's own best evaluator, an argmax over a half that is itself noisy, so on a panel whose best is unpredictable both rules miss it often and the difference sits in the loss part by construction; `scripts/decision/naming_loss.py` repeats the accounting against the corpus-best evaluator, the one with the highest payoff over every cluster, fixed before any split, on which a miss can have negative regret; on WMT23 that evaluator is GEMBA-MQM-src on zh-en, MetricX-23-refA on en-de and BERTscore-refB on he-en, where the limit choice of pooling is XCOMET-Ensemble-refA and XCOMET-Ensemble-refB and the grouped mean's mbr-metricx-qe-src and XCOMET-Ensemble-refB, so on those two pairs neither rule selects the corpus-best on more than 0% of draws: over the primary configurations at half the clusters the grouped mean selects it on 57% of draws against 41% and the split is +1.6 naming and +0.0 loss (+0.3 and +0.3 without HANNA relevance), at eight clusters +0.6 and +0.6; on WMT23 +1.5 and −2.1, with he-en +0.0 and −3.3; on RewardBench 2 +1.8 and +0.0 at half and −0.5 and −1.9 at eight.

The ordering advantage against the gain, over configurations and budgets. Read after the fact (`scripts/decision/ordering_by_budget.py` and

Table 34: The between-cluster share as a tie-break, on the same draws and bootstrap replicates as Table 5, paired differences in points; the configuration rows and the first block of aggregate rows under best of n, the second block under the pairwise choice. grp, P: the grouped mean and the pairwise accuracy computed on the labelled clusters, against pooling. A: the grouped mean's evaluator where it and pooling agree, and where they differ the one of the two whose own scores have the larger between-cluster share η_p^2 on the read clusters, no label read for the share; Alo: the other one. Against pooling the two split the grouped mean's gain into the draws on which its evaluator has the larger share (A$-r_T$) and those on which it has the smaller (Alo$-r_T$), and the two columns sum to grp$-r_T$ by construction. A$-$grp, the arbiter's gain over the rule it modifies, equals $-$(Alo$-r_T$); A$-$C is its gain over the coin, the expected payoff of a random choice between the two selected evaluators; A$-$P its gain over the pairwise accuracy. switch: the share of draws on which A leaves the grouped mean's evaluator (the disagreement rate is in Table 33). Configuration rows at half the clusters, under best of n; the aggregate rows carry the estimate ± 1.96 bootstrap standard deviations. The direction of A is the association of Table 33, read on these configurations, so this is a check that the association survives at the level of draws and not a pre-registered test.

Configuration	grp$-r_T$	P$-r_T$	A$-r_T$	Alo$-r_T$	A$-$grp	A$-$C	A$-$P	switch %
so762 accuracy	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0	0
so762 fluency	+2.9	+2.6	+2.9	+0.0	+0.0	+1.5	+0.4	0
so762 prosodic	+4.7	+4.2	+4.9	−0.2	+0.2	+2.6	+0.7	2
so762	+0.0	+0.0	−0.1	+0.1	−0.1	−0.1	−0.2	1
FRANK factuality	+6.4	+6.4	+6.4	+0.0	+0.0	+3.2	+0.0	0
HANNA coherence	+0.6	+0.1	−0.0	+0.6	−0.6	−0.3	−0.2	6
HANNA complexity	+4.8	+4.9	+4.8	−0.0	+0.0	+2.4	−0.1	0
HANNA empathy	−2.5	−2.3	−2.1	−0.4	+0.4	−0.9	+0.2	6
HANNA engagement	−2.3	−2.8	−1.9	−0.4	+0.4	−0.7	+0.9	8
HANNA relevance	+21.0	+21.5	+21.0	+0.0	+0.0	+10.5	−0.5	0
HANNA surprise	+0.3	+0.1	+0.3	+0.0	−0.0	+0.2	+0.2	4
RealSumm	+2.4	+2.2	+2.1	+0.3	−0.3	+0.9	−0.1	14
SummEval coherence	−7.3	−4.3	+0.0	−7.3	+7.3	+3.6	+4.3	54
SummEval consistency	+4.8	+4.8	+4.8	+0.0	−0.0	+2.4	−0.0	0
SummEval fluency	−6.3	+2.3	−1.4	−5.0	+5.0	+1.8	−3.7	34
SummEval relevance	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0	0
WMT22 ende	−0.0	−0.8	−1.2	+1.2	−1.2	−1.2	−0.4	36
WMT22 enru	+1.0	+1.4	+0.2	+0.8	−0.8	−0.3	−1.2	21
WMT22 zhen	+0.2	+0.2	+0.3	−0.2	+0.2	+0.3	+0.1	9
WMT23 ende	−3.2	−3.0	−0.1	−3.2	+3.2	+1.6	+3.0	91
WMT23 heen	−3.3	+0.0	+0.3	−3.5	+3.5	+1.9	+0.2	70
WMT23 zhen	+4.7	+1.8	+0.0	+4.6	−4.6	−2.3	−1.7	99
RB2 factuality	+0.8	+0.8	+0.9	−0.1	+0.1	+0.5	+0.1	1
RB2 precise if	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0	0
RB2 math	+7.5	+7.4	+7.4	+0.1	−0.1	+3.7	+0.0	2
RB2 safety	+1.5	+1.4	+1.5	+0.0	+0.0	+0.8	+0.1	0
RB2 focus	−0.6	−0.6	−0.6	+0.0	+0.0	−0.3	−0.0	0
Utility under the best of n								
Primary, 19, 8 clusters	+1.2 [+0.1,+2.2]	+1.7 [+0.7,+2.8]	+0.6 [−0.2,+1.4]	+0.5 [−0.1,+1.2]	−0.5 [−1.2,+0.1]	+0.1 [−0.5,+0.6]	−1.1 [−1.8,−0.4]	29
Primary, 19, half	+1.6 [−0.2,+3.5]	+2.1 [+0.3,+4.0]	+2.2 [+0.5,+3.8]	−0.5 [−1.1,+0.0]	+0.5 [−0.0,+1.1]	+1.4 [+0.5,+2.2]	+0.0 [−0.7,+0.7]	10
WMT23 MQM, 3, 8 clusters	−0.0 [−1.4,+1.3]	+0.2 [−1.1,+1.5]	+0.2 [−0.5,+0.9]	−0.2 [−1.3,+0.8]	+0.2 [−0.8,+1.3]	+0.2 [−0.4,+0.8]	+0.0 [−1.0,+1.0]	55
WMT23 MQM, 3, half	−0.6 [−3.6,+2.4]	−0.4 [−3.3,+2.5]	+0.1 [−0.7,+0.9]	−0.7 [−3.4,+2.0]	+0.7 [−2.0,+3.4]	+0.4 [−1.0,+1.7]	+0.5 [−2.1,+3.0]	87
RewardBench 2, 5, 8 clusters	−2.4 [−3.8,−1.0]	−3.0 [−4.4,−1.6]	−2.1 [−3.4,−0.7]	−0.3 [−0.6,−0.1]	+0.3 [+0.1,+0.6]	−0.9 [−1.5,−0.2]	+0.9 [+0.4,+1.4]	9
RewardBench 2, 5, half	+1.8 [+0.2,+3.5]	+1.8 [+0.2,+3.4]	+1.8 [+0.3,+3.4]	−0.0 [−0.2,+0.2]	+0.0 [−0.2,+0.2]	+0.9 [+0.2,+1.7]	+0.0 [−0.2,+0.3]	1
Utility under the pairwise choice								
Primary, 19, 8 clusters	+0.5 [+0.2,+0.8]	+0.7 [+0.4,+1.0]	+0.4 [+0.2,+0.7]	+0.1 [−0.1,+0.3]	−0.1 [−0.3,+0.1]	+0.2 [+0.0,+0.3]	−0.3 [−0.5,−0.1]	29
Primary, 19, half	+0.9 [+0.4,+1.4]	+1.0 [+0.5,+1.5]	+1.0 [+0.5,+1.4]	−0.0 [−0.2,+0.1]	+0.0 [−0.1,+0.2]	+0.5 [+0.2,+0.7]	−0.0 [−0.2,+0.2]	10
WMT23 MQM, 3, 8 clusters	−0.2 [−0.5,+0.1]	+0.1 [−0.2,+0.4]	+0.1 [−0.1,+0.3]	−0.3 [−0.5,−0.1]	+0.3 [+0.1,+0.5]	+0.2 [+0.1,+0.3]	−0.0 [−0.2,+0.2]	55
WMT23 MQM, 3, half	−0.3 [−1.0,+0.4]	+0.3 [−0.4,+0.9]	+0.0 [−0.1,+0.2]	−0.3 [−1.0,+0.3]	+0.3 [−0.3,+1.0]	+0.2 [−0.2,+0.5]	−0.2 [−0.8,+0.3]	87
RewardBench 2, 5, 8 clusters	−0.8 [−1.6,−0.1]	−1.1 [−1.8,−0.3]	−0.7 [−1.4,−0.0]	−0.1 [−0.2,+0.0]	+0.1 [−0.0,+0.2]	−0.3 [−0.6,+0.0]	+0.3 [+0.1,+0.6]	9
RewardBench 2, 5, half	+0.9 [+0.1,+1.7]	+0.9 [+0.1,+1.6]	+0.9 [+0.2,+1.6]	−0.0 [−0.1,+0.1]	+0.0 [−0.1,+0.1]	+0.4 [+0.1,+0.8]	+0.0 [−0.1,+0.1]	1

`ordering_by_budget_family.py`, from the released result files). For each of the 27 configurations of the exploratory set and the two validation sets, and at each budget of the grid, a rule's ordering advantage is pooling's selection error over evaluator pairs minus its own and its utility gain the paired utility difference of the evaluators the two select, both on the evaluation half, which is the same at every budget. The configurations of a benchmark share clusters, so the 27 points are not independent and every interval of this paragraph, over resampled configurations, is descriptive; resampling the benchmarks as units, WMT22 and WMT23 as one, is given beside the main contrasts.

What the 27 show together. Table 37 and Figure 8: for the grouped Spearman mean under best of n the correlation of the two across configurations is between 0.62 and 0.73 from two to sixteen labelled clusters, 0.68 $[+0.47, +0.82]$ at eight, 0.48 at 32 and 0.19 $[-0.16, +0.48]$ at half (Spearman's ρ 0.63, 0.53 and 0.20), and under the pairwise choice 0.72 $[+0.53, +0.87]$, 0.29 and 0.04 $[-0.29, +0.44]$ (ρ 0.69, 0.35 and 0.30), while the mean advantage grows at every step, from -0.6 to +5.2 points under best of n. The budgets are nested subsets of the same draws scored on the same evaluation halves, so the steps are one trajectory and are not counted.

Table 35: Ordering against choice, under best of n, same draws as Table 5; every entry of a set row is the mean over its configurations of the per-configuration value of Table 36, and the leave-one-out rows drop the configuration named. Advantage: pooling's selection error minus the grouped mean's, in points, over all decisive pairs and per pair on the pairs that involve the evaluation half's best evaluator (with best) and on the rest; on the $K-1$ pairs with the best evaluator a rule's error is that evaluator's mean rank less one over $K-1$, so the with-best entry is the displacement of the best evaluator's mean rank between the two rules, in points, and says nothing about whether either rule names it. Selects best: the share of draws on which the rule selects that evaluator. Loss: the mean regret of the selected evaluator when it is not the best, in payoff points. Utility: the paired difference of Table 5, with its interval from the same replicates; it splits exactly, per configuration, as $(h_G - h_T)(L_T + L_G)/2 + (1 - (h_T + h_G)/2)(L_T - L_G)$, the symmetric member of a family of exact splits (Appendix O), whose first term is printed as the naming part and second as the loss part; the parts are point estimates without an interval, the set rows average the parts, and their printed rates and losses do not multiply to them. The reference is the evaluation half's own best evaluator; against the corpus-best evaluator, Appendix O.

		Advantage, points			Selects best, %		Loss, points		Utility parts		
Set (cfg)	b	all pairs	with best	rest	r_T	grp	r_T	grp	naming	loss	Utility grp$-r_T$
Primary (19)	half	+5.0	+3.0	+5.4	33	44	8.8	8.0	+1.4	+0.2	+1.6 [−0.2, +3.5]
Primary leaving out HANNA relevance (18)	half	+4.7	+1.6	+5.2	35	41	8.0	6.9	+0.2	+0.4	+0.5 [−1.2, +2.3]
Primary (19)	8	+2.4	+1.1	+2.7	27	30	13.7	12.9	+0.5	+0.6	+1.2 [+0.1, +2.2]
Primary leaving out HANNA relevance (18)	8	+2.1	−0.2	+2.5	28	29	12.9	12.1	−0.1	+0.7	+0.6 [−0.3, +1.6]
WMT23 MQM (3)	half	+12.4	+19.5	+12.0	11	27	5.7	6.4	+0.5	−1.1	−0.6 [−3.6, +2.4]
RewardBench 2 (5)	half	+1.8	+0.5	+1.8	59	75	4.3	2.2	+0.8	+1.0	+1.8 [+0.2, +3.5]
RewardBench 2 leaving out math (4)	half	+2.7	+0.5	+2.7	70	81	2.5	1.6	+0.3	+0.1	+0.4 [−1.0, +1.8]
RewardBench 2 (5)	8	−3.8	−5.7	−3.8	22	14	18.4	20.7	−0.4	−2.0	−2.4 [−3.8, −1.0]
RB2 w/o allenai (5)	half	−0.3	+1.4	−0.4	67	77	2.4	2.1	+0.4	+0.3	+0.6 [−0.7, +2.0]
PRM800K (2)	8	−6.7	−13.1	−2.5	73	49	29.4	29.3	−7.2	−0.1	−7.3 [−9.6, −5.0]

Half is not one budget, and machine translation carries the last step. Half of a configuration's clusters is 48 to 249 clusters off machine translation, at most 64 on 15 of those 21 configurations, and 230 to 939 on the 6 of WMT22 and WMT23, which are one task in consecutive years with overlapping metric families and are treated as one family here; 32 is the last budget every configuration holds. On that budget the correlation at eight clusters less the one at 32 is +0.20 $[-0.01, +0.40]$ under best of n (benchmarks resampled $[-0.02, +0.37]$; +0.09 to +0.26 leaving one of the five task families out, clear of zero in 1) and +0.43 $[+0.17, +0.67]$ under the pairwise choice ($[+0.20, +0.69]$; +0.38 to +0.46, clear of zero in all 5). From eight clusters to half it is +0.49 $[+0.18, +0.82]$ and +0.68 $[+0.26, +1.06]$ on all configurations, and without machine translation +0.22 $[-0.06, +0.54]$, positive on 94% of resamples, and +0.51 $[-0.04, +0.93]$; leaving one task family out it runs from +0.22, the value without translation, to +0.66, clear of zero without any of the other four. The machine-translation configurations go from a mean advantage of +2.3 points and a mean gain of +1.2 at eight segments to +8.7 and +2.0 at 32 and +14.1 (9 to 19) and -0.1 (-3.3 to +4.7, positive on 3 of 6) at half, where the grouped mean names the evaluation half's best metric on 30% of draws against 18% for pooling and its regret is 4.1 points against 4.0: the grouped mean names the best metric more often and its regret is no smaller. With the corpus bootstrap of the family mean, the six configurations' gain at half is -0.1 $[-1.9, +1.7]$ against +1.8 $[+0.1, +3.6]$ for the other 21, a difference of -1.9 $[-4.5, +0.6]$; under the pairwise choice -0.1 $[-0.5, +0.3]$ against +1.1 $[+0.6, +1.5]$, a difference of -1.2 $[-1.8, -0.6]$. At eight segments the same six gain +1.2 $[+0.3, +2.1]$ under best of n. The least-squares line of gain on advantage over the other 21 at half, $+0.33 + 0.56\times$ advantage, gives about eight points of gain at machine translation's mean advantage; that line, fixed, is the yardstick of the third protocol (Appendix T). The other 21 go from +0.7 and +0.1 to +1.7 and +1.3 and to +2.7 and +1.8, with correlations of 0.73, 0.60 and 0.51 (ρ 0.69, 0.64 and 0.55) and a slope of gain on advantage of 0.70, 0.63 and 0.56 points of recovery per point of error, against 0.62, 0.41 and 0.15 with machine translation in; under the pairwise choice 0.75, 0.36 and 0.24 (ρ 0.72, 0.57 and 0.54), slopes 0.20, 0.09 and 0.06.

Other rules, and rule pairs without pooling. The pairwise accuracy on the labelled clusters behaves as the grouped mean does (0.73, 0.49 and 0.14 under best of n; under the pairwise choice a fall to 32 clusters of +0.45 $[+0.19, +0.67]$, clear of zero with any family left out), which is one finding and not two, the two statistics ordering alike. The within-cluster Pearson coefficient shows less: 0.55 to 0.35 on all configurations, a difference of +0.19 $[-0.11, +0.41]$, and 0.57 to 0.50 off machine

Table 36: Table 35 by configuration, under best of n, same draws as Table 5: the grouped mean's ordering advantage over pooling in points over all decisive pairs and per pair on the pairs with the evaluation half's best evaluator (the displacement of that evaluator's mean rank) and on the rest, the share of draws on which each rule selects the best evaluator, the mean loss given a miss, and the utility difference with its naming and loss parts under the symmetric split (defined in Table 35); the rows at half the clusters first, then the primary and RewardBench 2 rows at eight and PRM800K at eight problems.

			Advantage, pts			Selects best, %		Loss, pts		Parts		
Configuration	b	Ev.	all	with best	rest	r_T	grp	r_T	grp	naming	loss	Utility
speechocean762 accuracy	half	9	−0.3	−2.6	+0.4	83	83	3.8	3.8	+0.0	+0.0	+0.0
speechocean762 fluency	half	9	+4.8	+3.8	+5.2	9	20	11.6	9.5	+1.2	+1.8	+2.9
speechocean762 prosody	half	9	+8.6	+10.5	+8.0	29	57	11.3	7.6	+2.6	+2.1	+4.7
speechocean762 total	half	9	+1.9	−5.4	+4.0	62	62	4.9	4.8	+0.0	+0.0	+0.0
FRANK factuality	half	11	+11.2	+3.1	+13.0	65	97	18.5	1.6	+3.2	+3.2	+6.4
HANNA coherence	half	19	+0.1	−1.6	+0.3	20	26	7.7	7.5	+0.4	+0.2	+0.6
HANNA complexity	half	19	+4.7	+3.7	+4.8	10	36	9.0	5.2	+1.8	+3.0	+4.8
HANNA empathy	half	19	−1.7	−4.5	−1.3	17	13	8.1	10.6	−0.3	−2.1	−2.5
HANNA engagement	half	19	+0.1	−7.7	+1.1	22	7	10.8	11.5	−1.7	−0.6	−2.3
HANNA relevance	half	19	+10.3	+29.1	+8.1	0	92	23.2	27.4	+23.3	−2.3	+21.0
HANNA surprise	half	19	−0.8	−4.6	−0.4	2	2	14.8	14.6	+0.0	+0.3	+0.3
RealSumm LitePyramid	half	16	+16.1	+26.9	+14.5	11	30	6.9	5.5	+1.2	+1.2	+2.4
SummEval coherence	half	16	−0.5	−3.7	−0.1	92	41	4.9	13.0	−4.6	−2.7	−7.3
SummEval consistency	half	16	+0.7	+3.5	+0.3	10	76	6.0	2.6	+2.8	+1.9	+4.8
SummEval fluency	half	16	−0.1	−3.6	+0.4	30	0	13.7	15.9	−4.5	−1.9	−6.3
SummEval relevance	half	16	−6.8	+0.0	−7.8	100	100	0.0	0.0	+0.0	+0.0	+0.0
WMT22 en-de	half	18	+14.2	−1.8	+16.2	31	41	4.1	4.7	+0.4	−0.4	−0.0
WMT22 en-ru	half	17	+14.3	+2.7	+15.9	26	34	4.0	2.9	+0.3	+0.7	+1.0
WMT22 zh-en	half	18	+18.7	+9.5	+19.9	17	21	3.3	3.2	+0.1	+0.0	+0.2
WMT23 en-de	half	32	+12.3	+3.1	+12.9	20	3	4.8	7.3	−1.1	−2.2	−3.2
WMT23 he-en	half	32	+16.1	+32.5	+15.0	3	3	6.8	10.1	−0.0	−3.2	−3.3
WMT23 zh-en	half	32	+9.0	+22.8	+8.1	9	76	5.6	1.7	+2.4	+2.2	+4.7
RB2 factuality	half	174	+3.1	+0.1	+3.2	85	91	6.4	1.8	+0.3	+0.6	+0.8
RB2 precise IF	half	174	+3.0	+0.0	+3.1	100	100	0.0	0.0	+0.0	+0.0	+0.0
RB2 math	half	174	−1.8	+0.5	−1.9	18	53	11.7	4.5	+2.8	+4.7	+7.5
RB2 safety	half	174	+2.4	+2.3	+2.4	8	66	2.6	2.4	+1.4	+0.1	+1.5
RB2 focus	half	174	+2.1	−0.3	+2.2	86	67	0.9	2.3	−0.3	−0.3	−0.6
RB2 factuality (w/o allenai)	half	42	−0.1	−0.0	−0.1	99	99	1.2	1.2	+0.0	+0.0	+0.0
RB2 precise IF (w/o allenai)	half	42	−1.3	+0.0	−1.4	100	100	0.0	0.0	+0.0	+0.0	+0.0
RB2 math (w/o allenai)	half	42	−1.8	+0.9	−1.9	40	53	7.4	4.4	+0.7	+1.6	+2.4
RB2 safety (w/o allenai)	half	42	+1.4	+7.3	+1.1	8	66	2.6	2.4	+1.5	+0.1	+1.5
RB2 focus (w/o allenai)	half	42	+0.0	−1.3	+0.1	86	65	0.9	2.3	−0.3	−0.3	−0.7
speechocean762 accuracy	8	9	+1.6	+3.1	+1.2	53	63	10.6	12.6	+1.2	−0.8	+0.4
speechocean762 fluency	8	9	+1.6	−1.2	+2.4	17	15	9.5	9.4	−0.2	+0.1	−0.0
speechocean762 prosody	8	9	+1.6	−3.6	+3.1	40	26	9.4	11.5	−1.5	−1.4	−2.9
speechocean762 total	8	9	+3.7	−1.3	+5.2	38	42	13.3	10.0	+0.5	+2.0	+2.5
FRANK factuality	8	11	+5.1	−10.5	+8.6	43	43	22.3	19.1	+0.1	+1.8	+2.0
HANNA coherence	8	19	+1.3	−0.5	+1.5	13	15	14.6	12.3	+0.4	+2.0	+2.3
HANNA complexity	8	19	+2.9	+1.6	+3.0	13	22	11.6	9.1	+0.9	+2.0	+2.9
HANNA empathy	8	19	+0.2	−4.0	+0.6	17	11	10.5	12.7	−0.7	−2.0	−2.7
HANNA engagement	8	19	+0.9	−2.2	+1.2	11	11	13.5	12.2	+0.0	+1.1	+1.2
HANNA relevance	8	19	+7.3	+23.6	+5.3	9	50	27.0	27.9	+11.1	−0.6	+10.5
HANNA surprise	8	19	+2.9	−2.4	+3.5	13	9	16.1	15.4	−0.7	+0.6	−0.2
RealSumm LitePyramid	8	16	+10.2	+20.0	+8.8	11	25	9.0	6.3	+1.1	+2.2	+3.3
SummEval coherence	8	16	−1.5	−1.6	−1.5	63	44	12.9	13.0	−2.4	−0.0	−2.4
SummEval consistency	8	16	+0.0	+1.3	−0.1	21	55	6.5	6.0	+2.1	+0.3	+2.4
SummEval fluency	8	16	−0.1	−3.9	+0.4	36	16	13.7	15.3	−2.8	−1.2	−4.0
SummEval relevance	8	16	−2.9	+0.1	−3.4	83	81	24.3	24.0	−0.5	+0.0	−0.4
WMT22 en-de	8	18	+2.6	+0.9	+2.8	11	19	13.3	12.1	+1.0	+1.0	+2.0
WMT22 en-ru	8	17	+5.2	+1.4	+5.7	13	18	13.7	8.0	+0.5	+4.8	+5.3
WMT22 zh-en	8	18	+2.3	+0.2	+2.6	10	8	7.9	7.9	−0.1	+0.0	−0.1
RB2 factuality	8	174	−3.1	−5.7	−3.1	17	10	18.3	17.7	−1.2	+0.5	−0.7
RB2 precise IF	8	174	+0.5	−0.8	+0.5	29	35	42.7	40.7	+2.6	+1.3	+4.0
RB2 math	8	174	−4.0	−6.7	−4.0	12	11	18.5	20.6	−0.2	−1.8	−2.0
RB2 safety	8	174	−7.4	−6.6	−7.4	9	8	7.6	12.9	−0.1	−4.8	−4.9
RB2 focus	8	174	−4.9	−8.5	−4.8	42	5	5.0	11.6	−3.1	−5.1	−8.2
PRM800K answer correct problem	8	5	−4.5	−9.4	−1.3	86	59	33.5	32.5	−8.7	+0.3	−8.4
PRM800K human progress problem	8	5	−9.0	−16.8	−3.7	61	38	25.2	26.2	−5.7	−0.5	−6.2

translation (+0.08 $[-0.24, +0.34]$), with a mean advantage that grows from +1.6 to +2.7 points and off machine translation not at all (+1.3 to +1.1). The loosening between eight clusters and 32 is not particular to a grouped statistic against pooling. For best of n computed on the labelled clusters against pooling, a rule that orders worse than pooling at eight clusters, the correlation of its advantage

with its gain is 0.73, 0.47 and 0.21 at eight, 32 and half (0.77, 0.52 and 0.41 off machine translation); for the grouped mean against the within-cluster coefficient, two grouped statistics, 0.56, 0.38 and 0.29 (0.66, 0.49 and 0.58); for best of n on the labelled clusters against the grouped mean it does not fall, 0.57, 0.46 and 0.59. The fall of the grouped mean against pooling from eight clusters to 32 exceeds these pairs' by -0.07 to +0.10 under best of n and by +0.24 to +0.33 under the pairwise choice, no difference clear of zero over resampled configurations (+0.33 $[-0.11, +0.84]$ against the two grouped statistics, +0.24 $[-0.05, +0.54]$ against best of n and pooling). So what the panels show between eight clusters and 32 is what two rules with different noise and different limits would show. Nor is machine translation at half particular to grouping against pooling: there the grouped mean orders +5.8 points better than the within-cluster coefficient and its metric is worth -0.8, the pairwise accuracy +6.0 and -0.8 against the same coefficient, and best of n on the labelled segments +17.9 and +0.3 against pooling. On these panels, with hundreds of labelled segments, an ordering advantage of one rule over another comes with no gain whichever two rules are compared.

Across configurations, not along the budget. The correlations above are across configurations at one budget. Along the budget of one configuration the advantage and the gain mostly rise together: their correlation over the ten budgets has a median of 0.54 and is positive on 21 of 27 configurations. Part of that is a trend the two share as both rules' noise falls (the two means correlate at 0.94 across budgets): the change in advantage from eight clusters to half correlates with the change in gain at 0.13 (0.21 under the pairwise choice), so the configurations whose advantage grew more did not gain more. On machine translation the two part along the budget as well: the six configurations' mean gain is +1.2 at eight segments, +2.0 at 32, +1.5 at 48, +1.2 at 64, +0.7 at 96 and +0.1 at 192, while their mean advantage climbs from +2.3 to +13.5 points. Across all 27, the mean gain per point of mean advantage is 0.34 at eight clusters, 0.45 at 32 and 0.27 at half (0.19, 0.75 and 0.68 off machine translation). By configuration the advantage is larger at half than at eight on 19 of 27 and not positive at half on 7. The two same-model panels are outside the 27: on PPE, six evaluators scored against a fixed evaluation set, the grouped mean's mean ordering advantage does not grow, +3.1 points at eight prompts and +3.3 at the 48 of half the pool, and its correlation with the gain across the five tasks stays between 0.85 and 0.97 from three prompts up, one task with a large advantage and gain and one with both negative carrying it; on PRM800K the advantage is negative at every budget under best of n; with the two PRM800K labels added to the 27 the correlations are 0.74 at eight and 0.20 at half; on RewardBench 2 without its authors' models, which shares RewardBench 2's clusters, the correlation across the five subsets is 0.82 at eight prompts and -0.15 at half with reversals between, five points each being too few to mean much (`ordering_by_budget_parts.py`). Separately on the 19 exploratory configurations the correlation goes from +0.67 at eight clusters to +0.37 at half and on the 8 of the two validation sets from +0.55 to -0.65, the second being the contrast between WMT23 and RewardBench 2.

Noise. A low correlation between two noisy quantities can be the noise, so the noise was measured at the two budgets with a corpus bootstrap: at half the gains have a variance of 27.6 squared points across configurations, of which the bootstrap standard deviations of the single configurations account for 11.5, a reliability of 0.58, and a gain exactly linear in the advantage, with that remaining variance and observed with each configuration's own bootstrap noise, correlates with it at 0.78 $[0.58, 0.89]$ over 20,000 simulations and at 0.19 or less on none of them; at eight the reliability is 0.75 and the same construction gives 0.88 $[0.74, 0.94]$, above the observed 0.68 on all but 0.7% of simulations. Without machine translation the observed correlation sits at the same place under its ceiling at both budgets, 0.73 against 0.88 $[0.73, 0.95]$ at eight (at or below the observed on 2.4% of simulations) and 0.51 against 0.80 $[0.51, 0.92]$ at half (2.4%) under best of n, and lower at half under the pairwise choice (0.24 against 0.75 $[0.43, 0.90]$, 0.2%): the relation is loose at every budget, and off machine translation under best of n no looser at half than at eight. That simulation leaves the advantage's own noise out. A second one (`r15_thesis_checks.py`) observes both quantities with noise: the true advantages are the observed ones shrunk to the variance left after their bootstrap noise (reliability 0.80 at eight clusters and 0.84 at half), the true gains are exactly linear in them with the variance left after theirs, and the two errors of a configuration are drawn jointly, from the same replicate row, so that their correlation within a configuration (0.45 at eight, 0.23 at half on average) is kept. The reference barely moves, because the two errors being positively correlated offsets the advantage's noise: 0.88 $[0.75, 0.94]$ at eight and 0.79 $[0.59, 0.90]$ at half on all 27 under best of n (at or below the observed on 0.2% and below 0.1% of draws), 0.87 and 0.78 without machine translation (2.3% and 3.2%), and under the pairwise choice 0.88 and 0.76 (1.7% and below 0.1%), 0.89 and 0.74 without it (3.1% and 0.4%). Both simulations draw the noise independently across configurations, take the bootstrap's

standard deviations at face value and assume a linear truth, so their correlations are references and not bounds. The relation at small budgets is not the convention for an undefined statistic, which counts half an error and hands the choice to the draw's priority: no rule is undefined on more than 2.1% of pairs at any budget and on less than half a percent from 4 clusters on (Appendix J).

Where the advantage sits. The bins of Table 35 split the decisive pairs by whether they hold the evaluation half's best evaluator: those that do not are 89% of the pairs and carry 92% of the grouped mean's advantage at half. With eight clusters the utility gain correlates across configurations with the per-pair advantage on the pairs that hold the best evaluator at 0.64 and on the others at +0.60; at half at 0.47 and +0.12 (0.72 and +0.70, then 0.44 and -0.01, under the pairwise choice). Table 38 gives the same split at every budget of the grid. The entry for the pairs that hold the best evaluator is a displacement of that evaluator's rank, and the gain is decided by which evaluator is ranked first, so that part's relation to the gain is partly by construction and the information is in the rest. The rest's correlation with the gain falls by more than the top part's where the all-pairs fall is itself clear: under the pairwise choice from eight clusters to 32 by +0.33 $[+0.12, +0.57]$ more (+0.25 $[+0.05, +0.43]$ without machine translation), and under best of n from eight clusters to half by +0.31 $[+0.05, +0.60]$ more; under best of n from eight to 32 the contrast is +0.16 $[-0.05, +0.36]$ and without machine translation to half +0.21 $[-0.04, +0.52]$, neither clear of zero (configurations resampled). Off machine translation the part at the top stays at 0.72 at eight clusters, 0.74 at 32 and 0.72 at half under best of n, where the rest goes from 0.64 to 0.53 and 0.43, which is the point of Mathur et al. (2020) one level up, with its dependence on the label budget; on machine translation it is not that point, the advantage at half being as large on the pairs that hold the best metric, +11.5 points, as on the others, +14.7. A reading of the whole is that with few clusters a rule's errors come from the noise of its statistic, which reaches the top of a panel as it reaches the rest, and with many what remains is what two statistics measure differently, which pays only where the evaluators it reorders differ in payoff. The reading was not tested on the benchmarks. A simulation without clusters shows that it suffices for the pattern and that the pattern is not particular to grouping (`scripts/sim/ordering_gain_mechanism.py`): 27 panels of 8 to 20 evaluators, one rule whose statistic is the payoff plus noise shrinking with the square root of the labelled units, and another with its own noise and a fixed distortion of its limit ordering whose size varies by panel, 200 repetitions. When the distortion falls on evaluators away from the top three, the across-panel correlation of advantage and gain goes from 0.85 at eight units to 0.72 at 32 and 0.46 at 512, the part of the advantage on the pairs that hold the best evaluator keeps its relation to the gain (0.87, 0.79 and 0.82) and the rest loses it (0.82, 0.69 and 0.40), which is the shape of Table 38 off machine translation; when it falls on the top three the all-pairs correlation falls further (0.74 to 0.24) and the part at the top keeps less (0.67 to 0.48). So two rules with different noise and different limits loosen in this way whatever the rules are, and what the benchmarks add is where the limits differ: away from the top on most panels, and on machine translation among metrics too close in payoff for the choice to matter.

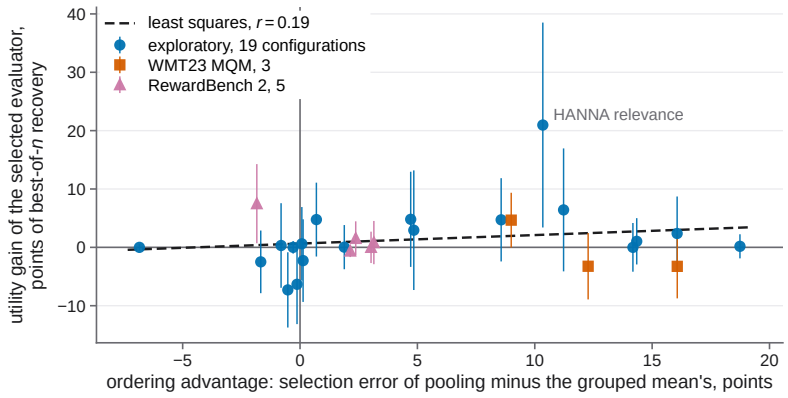


Figure 10: **Ordering advantage against utility gain, by configuration.** The grouped mean's ordering advantage over pooling (points of selection error over evaluator pairs) against the utility gain of the evaluator it selects (points of best-of-n recovery, estimate ± 1.96 bootstrap standard deviations), one point per configuration, half the clusters, 300 splits.

Table 37: The ordering advantage against the utility gain across configurations, by label budget, read after the fact (`scripts/decision/ordering_by_budget.py` and `ordering_by_budget_family.py`). A rule is computed on b clusters of the selection half ($C/2$: all of it, 48 to 249 clusters off machine translation and 230 to 939 on it, so not one budget; 32 is the last budget every configuration holds) and scored on the evaluation half, the same at every b; per configuration the advantage is pooling's selection error over evaluator pairs minus the rule's and the gain the paired utility difference of the selected evaluators, 300 splits. Columns for the grouped Spearman mean: configurations holding the budget, mean advantage, mean gain, least-squares slope of gain on advantage, Pearson r with the 2.5th to 97.5th percentiles over 10,000 resamples of configurations (descriptive: configurations of a benchmark share clusters) and Spearman's ρ, on all configurations; r, ρ, mean advantage and mean gain without the six machine-translation configurations (WMT22 and WMT23, one task family); the mean advantage and gain of those six; then r on all configurations for the within-cluster Pearson coefficient and the pairwise accuracy on the labelled clusters, each against pooling. Last rows of a block: differences of r between budgets, configurations resampled in pairs, on all configurations, without machine translation, and the range leaving one of the five task families out; below each, the interval with benchmarks resampled as units (seven units, WMT22 and WMT23 one; by Appendix Z an interval over so few units covers less than its nominal level) and the same contrast under the Spearman coefficient, which does not clear zero at any budget pair.

		grouped mean against pooling, all					without machine translation				MT	r_W	pair
b	cfg.	adv.	gain	slope	r	ρ	r	ρ	adv.	gain	adv. / gain	r	r
best of n, points of recovery													
2	27	−0.6	−0.9	0.79	0.68 [−0.05, +0.88]	0.32	0.71	0.34	−0.6	−1.0	−0.6 / −0.5	0.60	0.72
3	27	−0.2	−0.7	0.72	0.66 [+0.15, +0.85]	0.51	0.68	0.51	−0.3	−0.8	−0.2 / −0.4	0.60	0.65
4	27	+0.2	−0.4	0.72	0.73 [+0.42, +0.87]	0.66	0.76	0.65	+0.1	−0.5	+0.4 / +0.1	0.57	0.74
6	27	+0.7	−0.1	0.66	0.70 [+0.45, +0.84]	0.66	0.77	0.76	+0.5	−0.4	+1.3 / +1.0	0.59	0.76
8	27	+1.1	+0.4	0.62	0.68 [+0.47, +0.82]	0.63	0.73	0.69	+0.7	+0.1	+2.3 / +1.2	0.55	0.73
12	27	+1.7	+0.6	0.54	0.62 [+0.42, +0.78]	0.65	0.66	0.71	+1.0	+0.3	+4.0 / +1.7	0.56	0.69
16	27	+2.1	+0.8	0.53	0.62 [+0.45, +0.77]	0.68	0.68	0.69	+1.2	+0.5	+5.4 / +1.9	0.54	0.68
24	27	+2.8	+1.1	0.43	0.51 [+0.28, +0.70]	0.54	0.62	0.63	+1.6	+1.0	+7.3 / +1.7	0.50	0.55
32	27	+3.3	+1.5	0.41	0.48 [+0.24, +0.68]	0.53	0.60	0.64	+1.7	+1.3	+8.7 / +2.0	0.49	0.49
48	21	+4.4	+1.0	0.16	0.28 [−0.10, +0.55]	0.33						0.44	0.26
$C/2$	27	+5.2	+1.4	0.15	0.19 [−0.16, +0.48]	0.20	0.51	0.55	+2.7	+1.8	+14.1 / −0.1	0.35	0.14
r at eight clusters less r at 32					+0.20 [−0.01, +0.40]		+0.13 [−0.08, +0.40]		a family out: +0.09 to +0.26			+0.05	+0.24
the same, benchmarks resampled as units					[−0.02, +0.37]		[−0.07, +0.39]						
the same contrast in ranks (Spearman)					+0.10 [−0.22, +0.42]		+0.05 [−0.37, +0.46]						
r at eight clusters less r at $C/2$					+0.49 [+0.18, +0.81]		+0.22 [−0.06, +0.54]		a family out: +0.22 to +0.66			+0.19	+0.59
the same, benchmarks resampled as units					[+0.04, +1.00]		[−0.01, +0.67]						
the same contrast in ranks (Spearman)					+0.43 [−0.02, +0.84]		+0.15 [−0.32, +0.62]						
pairwise choice, points of accuracy													
2	27	−1.3	−0.4	0.29	0.69 [+0.23, +0.87]	0.57	0.71	0.51	−1.1	−0.4	−1.9 / −0.3	0.82	0.69
3	27	−1.0	−0.3	0.25	0.64 [+0.28, +0.85]	0.63	0.67	0.63	−0.8	−0.3	−1.8 / −0.2	0.80	0.66
4	27	−0.4	−0.2	0.23	0.76 [+0.62, +0.87]	0.81	0.78	0.83	−0.1	−0.2	−1.5 / −0.1	0.80	0.76
6	27	+0.4	−0.0	0.20	0.76 [+0.64, +0.88]	0.75	0.81	0.84	+0.8	−0.0	−1.0 / +0.1	0.76	0.77
8	27	+0.9	+0.2	0.19	0.72 [+0.53, +0.87]	0.69	0.75	0.72	+1.3	+0.2	−0.3 / +0.1	0.62	0.74
12	27	+1.9	+0.4	0.15	0.60 [+0.36, +0.82]	0.67	0.62	0.71	+2.2	+0.4	+0.8 / +0.2	0.51	0.62
16	27	+2.7	+0.5	0.13	0.55 [+0.29, +0.78]	0.54	0.57	0.62	+2.9	+0.5	+2.1 / +0.3	0.36	0.56
24	27	+3.7	+0.6	0.09	0.40 [+0.14, +0.66]	0.45	0.45	0.59	+3.7	+0.7	+3.8 / +0.2	0.35	0.40
32	27	+4.4	+0.8	0.07	0.29 [+0.05, +0.57]	0.35	0.36	0.57	+4.3	+0.9	+4.9 / +0.3	0.32	0.28
48	21	+5.0	+0.7	0.01	0.07 [−0.26, +0.52]	0.15						0.25	0.08
$C/2$	27	+7.5	+0.8	0.01	0.04 [−0.29, +0.44]	0.30	0.24	0.54	+5.8	+1.1	+13.3 / −0.1	0.41	0.03
r at eight clusters less r at 32					+0.43 [+0.17, +0.67]		+0.39 [+0.09, +0.62]		a family out: +0.38 to +0.46			+0.31	+0.45
the same, benchmarks resampled as units					[+0.20, +0.69]		[+0.13, +0.76]						
the same contrast in ranks (Spearman)					+0.34 [−0.05, +0.74]		+0.15 [−0.24, +0.54]						
r at eight clusters less r at $C/2$					+0.68 [+0.26, +1.06]		+0.51 [−0.04, +0.93]		a family out: +0.51 to +0.87			+0.21	+0.71
the same, benchmarks resampled as units					[+0.16, +0.93]		[+0.00, +0.85]						
the same contrast in ranks (Spearman)					+0.39 [−0.14, +0.89]		+0.18 [−0.33, +0.77]						

The scale of the gain. Weighting the six benchmarks equally rather than the configurations, the primary gain is +2.1 points $[-0.3, +4.4]$ under best of n and +1.1 $[+0.3, +2.0]$ under the pairwise choice, the interval from the benchmark groups' independent bootstrap streams. `scripts/decision/oracle_gap.py` pays two references on the main run's own draws beside the selected evaluators: the evaluation half's own best evaluator, the "best" column of Table 5, an oracle read off that half; and the corpus-best evaluator, the one with the highest payoff over every cluster of the corpus, fixed before any split and scored on the evaluation half like the selected ones, which earns 51.6% recovery and 71.0% pairwise accuracy there over the primary configurations against 53.2% and 71.1% for the half's own best. The grouped mean's evaluator closes 25% and 62% of the distance from pooling's

Table 38: Where the ordering advantage sits, by label budget, read after the fact (`top_pairs.py` at every budget of the grid and `ordering_by_budget_parts.py`; 27 configurations, 300 splits). The decisive evaluator pairs of a split are those that hold the evaluation half's best evaluator (top) and the others (rest, nine pairs in ten); per configuration the grouped mean's advantage over pooling is its error difference on each kind of pair, in points. Columns: the mean advantage on the two kinds under best of n; the correlation across configurations of each with the utility gain, on all configurations and without the six of machine translation; the mean advantage on the two kinds on machine translation; and the same correlations under the pairwise choice. The entry for the top pairs is a displacement of the best evaluator's rank, so its relation to the gain is partly by construction and the information is in the rest, whose correlation with the gain falls by more than the top part's where the all-pairs fall is clear (the pairwise choice from eight clusters to 32, best of n from eight to half; contrasts with intervals in the text of this appendix). On machine translation the advantage at the top is large as well, and comes with no gain (Table 37).

	best of n								pairwise choice			
	mean adv.		r with gain, all		r, without MT		mean adv., MT		r with gain, all		r, without MT	
b	top	rest	top	rest	top	rest	top	rest	top	rest	top	rest
2	−2.0	−0.5	0.66	0.64	0.67	0.67	−0.1	−0.7	0.64	0.69	0.67	0.71
3	−1.8	−0.1	0.65	0.58	0.69	0.60	−0.2	−0.1	0.63	0.63	0.67	0.65
4	−0.9	+0.4	0.71	0.65	0.75	0.68	+1.3	+0.3	0.72	0.75	0.76	0.77
6	−0.2	+0.9	0.74	0.60	0.78	0.68	+2.3	+1.3	0.72	0.75	0.79	0.80
8	+0.3	+1.3	0.64	0.60	0.72	0.64	+3.2	+2.3	0.72	0.70	0.75	0.73
12	+1.0	+1.9	0.59	0.56	0.69	0.59	+4.3	+4.1	0.63	0.56	0.65	0.58
16	+1.7	+2.3	0.63	0.56	0.74	0.61	+5.7	+5.5	0.63	0.49	0.65	0.52
24	+2.5	+3.0	0.63	0.45	0.76	0.55	+7.2	+7.5	0.62	0.32	0.64	0.39
32	+3.2	+3.4	0.62	0.41	0.74	0.53	+8.7	+8.9	0.56	0.21	0.57	0.29
$C/2$	+4.4	+5.5	0.47	0.12	0.72	0.43	+11.5	+14.7	0.44	−0.01	0.44	0.20

evaluator to the first under the two payoffs and 33% and 67% of the distance to the second; against the second the within-cluster coefficient's evaluator closes 25% and 41%, and at eight clusters the grouped mean's 14% and 19%.

A label-free diagnostic, read after the fact. The utility difference factors exactly into the share of draws on which two rules select different evaluators times the difference given that they do: over the 26 primary, PRM800K and PPE configurations of Table 32 the grouped mean and pooling differ on 45% of draws at half the clusters under best of n, and the grouped mean's evaluator is the better one on 61% of those draws (Appendix O).

The sign of the difference given disagreement goes with a quantity that needs no labels. Equation 1 credits an evaluator $\eta_p \eta_y r_B$ for ranking the clusters and discounts its within-cluster validity by $\sqrt{1-\eta_p^2}$, η_p^2 the between-cluster share of the variance of its own scores, so pooling can pass over an evaluator whose scores vary across clusters more than the labels do whatever it decides within clusters; the decomposition speaks to that direction only. On the 14 of the 27 primary, WMT23 and RewardBench 2 configurations on which the two rules select different evaluators when they use every cluster, the grouped mean's evaluator has the larger share where the grouped mean gains and the smaller where it loses on 12, 11 of the 13 whose gain has a readable sign at the printed precision (binomial $p = 0.022$ treating the configurations as independent, which those of one benchmark are not; Table 33). The two exceptions are HANNA engagement and WMT23 zh-en, the one WMT23 pair on which the grouped mean does select the corpus-best metric, where the share points the other way and the tie-break gives the whole gain back. Fixed as a rule after the association was read, on dependent configurations, and scored on the same draws, the share adds +0.5 points $[-0.0, +1.1]$ to the grouped mean at half the clusters, all of it from SummEval coherence and fluency, and loses or is level at the 9 smaller budgets, −0.5 at eight and within a tenth of a point of zero at 24 and 32 (Appendix O).

The share as a tie-break. `scripts/decision/etap_rule.py` pays the rules built from the between-cluster share on the main run's own draws and bootstrap replicates, replayed from its streams (`replay.py`), so that the grouped-mean-against-pooling column of Table 34 is Table 5's to the last digit: the arbiter A of the paragraph above, which takes the grouped mean's evaluator and, where pooling names another, the one of the two with the larger share η_p^2 of its own scores on the read

clusters; the coin C, the expected payoff of choosing one of the two selected evaluators at random, $(U_M(\text{grp}) + U_M(r_T))/2$ on every draw; the pairwise accuracy computed on the labelled clusters, P; and the pooled coefficient with its discount removed, $r_T/\sqrt{1-\eta_p^2}$ on the read clusters, ranking the whole panel, which is what Equation 1 would suggest computing. The smaller-share choice is printed too and is not a control: on every disagreeing draw it takes the evaluator the arbiter leaves, so its payoff against the grouped mean is minus the arbiter's against pooling by construction, which the table shows row for row. The comparisons that carry information are A against grp and A against P; A against C is the first plus half the grouped mean's gain over pooling, and A against r_T is that gain plus the first. The discount-free coefficient selects far worse evaluators, -7.3 points of recovery against pooling's over the primary configurations at half the clusters and -8.9 against the grouped mean's, most of it on the WMT22 pairs, because dividing by the discount inflates the between-cluster credit $\eta_p\eta_y r_B$ of an evaluator whose scores vary almost only across clusters; the discount is not separable from the credit, and a share is a diagnostic of a disagreement between two selected evaluators, not a correction to apply to a coefficient. The arbiter's gain over the grouped mean under best of n is -0.1, -0.1, -0.3, -0.4, -0.5, -0.3, -0.2, -0.1, -0.0 and +0.5 points at two, three, four, six, eight, twelve, sixteen, 24 and 32 clusters and at half, positive at half on 6 configurations and negative on 5; over the coin it is +0.1 at eight clusters and +1.4 at half, and against the pairwise accuracy +0.0 at half. It leaves the grouped mean's evaluator on 10% of draws at half, pooled over draws, and on 26% of the disagreeing draws averaged over configurations; under the pairwise choice at half it adds +0.0 to the grouped mean and +0.5 over the coin. A two-cluster construction shows what the share cannot carry (`scripts/sim/two_cluster_counterexample.py`): labels $0, 1, 2, 3$ in each of two clusters of four candidates; one evaluator scores them $2, 1, 0, 3$, the other $0, 1, 3, 2$, so the first has $r_W = 0.2$ and recovers all of the best-of-n gain and the second $r_W = 0.8$ and recovers one third, and their pairwise accuracies are 0.50 and 0.83; a constant added to either evaluator's scores in the second cluster leaves every one of those numbers where it is and moves that evaluator's share from 0 to 0.93 and its pooled coefficient with it, so the gap of the two shares runs from -0.93 to +0.93 over a grid of such offsets while no payoff moves. Nothing in the decomposition orders the evaluators' payoffs, which is why the share is offered as read on these panels. On RewardBench 2, where the labels have no between-prompt part, the pooled and within-cluster coefficients rank the models at a Spearman correlation of 0.921 to 0.996 over the five subsets and select the same top model on all but 2 of them, all but 1 without the authors' models, the models' own between-prompt shares running from 0.15 to 0.95 (`mechanism_etap.py`).

P What Pooling Leaves and What a Set Can Detect

Read after every protocol and graded by none. `scripts/decision/headroom_power_r23.py` reads, from the runs the paper already reports, pooling's regret on every set of Table 1, the mean over configurations of the points of best-of-n recovery by which pooling's pick falls short of the evaluation half's best evaluator on the same draw, beside the grouped mean's gain and its interval, and computes from the standard deviation s of the set mean's corpus bootstrap the gain a two-sided 5% test detects with 80% power, $2.8\,s$, and the power of that test against a gain as large as pooling's regret (Table 39). Three cautions go with it. The gain of a rule is pooling's regret less the rule's on every draw, so the regret bounds it; the bound is loose, the evaluation half's best being named with that half's noise, which no rule computed on the selection half can share. The power uses the bootstrap as the sampling distribution of the set mean, whose coverage Appendix AA measures on the exploratory and first protocol's sets only, and which it finds conservative with eight clusters and nominal to short at half (0.95 times the true standard deviation on the exploratory mean). And no share of the regret was fixed in advance as the gain worth detecting, so the closed shares of Table 39 and their intervals, which divide the gain's by the regret's estimate and do not carry the regret's own error, are bounds a reader may set against a threshold of their own, not tests of equivalence.

The intervals are the estimate plus or minus 1.96 bootstrap standard deviations. With the percentile interval of the same replicates in their place, seven of the 56 comparisons of the grouped mean against pooling on these sets, at both budgets and under both payoffs, change whether they exclude zero; at half under best of n one graded set clears zero under either interval, RewardBench 2 under the normal one and WMT24 ESA under the percentile one, each by less than half a point, and the percentile

Table 39: What pooling leaves and what each set can detect: the grouped mean against pooling under best of n, read after every protocol and graded by none. Reg.: the points of best-of-n recovery by which pooling's pick falls short of the evaluation half's best evaluator on the same draw, the mean over the set's configurations; it bounds any rule's gain and overstates what a rule computed on the selection half can reach. Gain: the grouped mean's, with its interval as the run reports it. Closed: the gain as a percentage of pooling's regret, its interval the gain's divided by the regret's point estimate, whose own sampling error is not carried. Det.: the gain a two-sided 5% test on the set mean detects with 80% power, 2.8 standard deviations of the corpus bootstrap. `scripts/decision/headroom_power_r23.py`.

	Eight labelled clusters				Half the clusters			
Set	Reg.	Gain	Closed (%)	Det.	Reg.	Gain	Closed (%)	Det.
exploratory	9.6	+1.2 [+0.1, +2.2]	+12 [+2, +23]	1.4	6.6	+1.6 [−0.2, +3.5]	+25 [−4, +53]	2.7
WMT23 MQM	9.9	+0.0 [−1.4, +1.3]	+0 [−14, +13]	1.9	5.2	−0.6 [−3.6, +2.4]	−12 [−69, +46]	4.2
RewardBench 2	14.4	−2.4 [−3.8, −1.0]	−16 [−26, −7]	2.0	2.6	+1.8 [+0.2, +3.5]	+70 [+9, +132]	2.3
RewardBench 2 without the authors' models	10.8	−1.7 [−3.2, −0.2]	−16 [−29, −2]	2.1	1.4	+0.6 [−0.7, +2.0]	+46 [−48, +141]	1.9
WMT24 MQM	16.6	+1.4 [−4.2, +6.9]	+8 [−25, +42]	7.9	5.6	−2.8 [−9.2, +3.7]	−49 [−164, +65]	9.2
WMT24 ESA	9.9	−0.1 [−1.8, +1.7]	−1 [−18, +17]	2.5	6.5	+2.0 [−0.3, +4.3]	+30 [−4, +65]	3.3
WMT25 MQM, primary metrics	18.1	+2.0 [−5.3, +9.2]	+11 [−29, +51]	10.4	13.2	+4.7 [−4.0, +13.4]	+35 [−30, +101]	12.5
WMT25 ESA, primary metrics	12.7	−0.4 [−2.1, +1.4]	−3 [−17, +11]	2.5	8.3	+1.5 [−0.8, +3.9]	+18 [−10, +46]	3.4
WMT25 MQM, LLM judges	17.9	+6.4 [+1.0, +11.7]	+36 [+6, +66]	7.7	6.3	+0.5 [−8.1, +9.1]	+8 [−128, +143]	12.3
WMT25 ESA, LLM judges	9.1	−0.1 [−1.4, +1.3]	−1 [−16, +15]	2.0	4.2	+1.1 [−0.5, +2.7]	+27 [−12, +65]	2.3
rerun after the grading with every rated WMT25 system								
WMT25 MQM, primary metrics, every system	18.9	+7.0 [+0.9, +13.1]	+37 [+5, +69]	8.7	11.3	+5.1 [−4.6, +14.7]	+45 [−41, +131]	13.8
WMT25 ESA, primary metrics, every system	13.8	+2.8 [+0.5, +5.1]	+21 [+4, +37]	3.3	10.2	+3.1 [+0.8, +5.5]	+31 [+8, +53]	3.3
WMT25 MQM, LLM judges, every system	17.2	+8.3 [+2.3, +14.4]	+48 [+13, +84]	8.6	3.4	−1.2 [−7.7, +5.3]	−36 [−231, +159]	9.3
WMT25 ESA, LLM judges, every system	8.0	+2.2 [+1.0, +3.3]	+27 [+12, +42]	1.7	5.2	+3.4 [+2.0, +4.8]	+65 [+39, +92]	2.0

interval sits on the replicates' mean, up to 4.5 points from the estimate on WMT25 MQM with the judges and every system kept, where the replicates are skewed.

Q Which Rule Loses to Pooling

Read after every protocol and graded by none. `scripts/decision/rule_safety_r22.py` reads, from the runs the paper already reports, every rule's paired utility difference against pooling with eight labelled clusters and at half on every set, and each configuration's where the run keeps its bootstrap; it recomputes nothing, so each interval keeps its run's convention (Table 41), and Table 40 counts them. The count is of descriptive intervals, on sets that share benchmarks, taken after the fact. What it adds to Table 12, on whose 58 cells, the exploratory set's, the first protocol's and PRM800K's, the within-cluster coefficient's interval never lies below zero, is the 82 configuration cells of the third and fourth protocols' sets, on which it never does either, and PPE, where with eight labelled prompts it gains +1.6 points with the reward models and +0.0 with the seven judges. Its interval lies above zero on 2 set means and 5 cells. The tally is of best of n at eight labelled clusters and at half. On the 222 further cells the runs keep a bootstrap for, the pairwise choice at both budgets and best of n at 32 labelled segments on WMT24 and WMT25, its interval lies below zero on 3: under the pairwise choice at half on speechocean762 prosody and WMT23 en-de, by about one point, and at 32 segments on WMT25 ESA en-bho; the grouped mean's on 8, the pairwise accuracy's on 4, the benchmark's own statistic's on 8 of 218 and best of n on the labels on 26. With WMT25's thirteen LLM judges and eight labelled segments the pairwise accuracy and the benchmark's own statistic gain where it does not, and with eight labelled clusters of two candidates each it picks worse than pooling on the exploratory configurations, though not at half (Appendix N). The WMT24 and WMT25 intervals keep 40 splits a replicate, which Appendix AA finds conservative at eight clusters.

How often a rule departs from pooling decides how often its interval can fall below zero. r_W names a different evaluator from pooling on 47% of the exploratory draws with eight labelled clusters and 38% at half, the grouped mean on 63% and 52%, and of the 140 configuration cells of Table 40 r_W and pooling disagree on fewer than one draw in ten on 19, the median cell on 44%. Per disagreement, the set mean's difference over its disagreement rate, r_W's pick gained +3.5 points on the exploratory set with eight labelled clusters and +3.2 at half, the grouped mean's +1.9 and +3.1; on WMT24 MQM at half, its worst set, r_W's difference per disagreement was −7.3 points and the grouped mean's −3.1. The coverage study of Appendix AA does not include r_W against pooling (Table 74), so these intervals carry no measured coverage.

Table 40: **Losses and wins against pooling, by rule.** Best of n (PPE: best of 16) with eight labelled clusters and with half, read after every protocol and graded by none. Set means: the 24 means of the first 12 sets of Table 41, each at both budgets; configurations: the 140 configuration cells of those runs that keep a bootstrap, 58 of them those of Table 12; below and above: how many intervals lie wholly below or above zero (own: of 20 and 136). Worst: the lowest set-mean estimate, in points of recovery. Advantage: the points of pooling's selection error the rule removes at half on the exploratory set.

	Set means		Configurations			
Rule against pooling	below	above	below	above	Worst	Advantage
r_W, within-cluster Pearson	0	2	0	5	−2.7	+2.2
grp, grouped Spearman mean	2	4	7	5	−7.3	+5.0
pairwise accuracy in clusters	2	6	5	7	−6.7	+5.0
own, the benchmark's statistic	1	3	5	13	−5.0	+2.2
best of n on the labels	4	0	13	2	−6.0	+2.2

Table 41: Every set mean behind Table 40: each rule's pick against pooling's, the paired utility difference in points of best-of-n recovery (PPE: best-of-16 accuracy), the estimate with its interval as each run reports it (300 splits a replicate on the exploratory set, the first protocol's sets and PRM800K, the frozen 40 on WMT24 and WMT25; PPE's reward models with the pool and the measurement set resampled; the fifth protocol's bootstrap for its judges). The last four rows of each budget keep every rated WMT25 system and are not counted in Table 40. Read after every protocol, graded by none; `scripts/decision/rule_safety_r22.py`.

Set	r_W	grp	pairwise acc.	own	best of n on labels
eight labelled clusters (PPE: prompts)					
exploratory	+1.6 [+0.7, +2.6]	+1.2 [+0.1, +2.2]	+1.7 [+0.7, +2.8]	+1.0 [+0.1, +1.8]	−4.4 [−6.0, −2.9]
WMT23 MQM	−0.1 [−1.1, +0.9]	+0.0 [−1.4, +1.3]	+0.2 [−1.1, +1.5]	+0.0 [−1.2, +1.2]	−3.2 [−4.7, −1.8]
RewardBench 2	+0.9 [−0.2, +2.1]	−2.4 [−3.8, −1.0]	−3.0 [−4.4, −1.6]	−3.4 [−4.9, −2.0]	−3.4 [−4.9, −2.0]
PRM800K	+0.0 [−1.4, +1.5]	−7.3 [−9.6, −5.0]	−6.7 [−8.9, −4.5]	–	−4.8 [−7.1, −2.6]
WMT24 MQM	−0.1 [−3.8, +3.7]	+1.4 [−4.2, +6.9]	+2.2 [−3.1, +7.6]	+1.6 [−3.9, +7.1]	−2.9 [−8.6, +2.8]
WMT24 ESA	+0.0 [−1.3, +1.3]	−0.1 [−1.8, +1.7]	+0.6 [−1.3, +2.5]	+1.0 [−1.0, +2.9]	−1.0 [−2.9, +0.9]
WMT25 MQM, primary metrics	+1.6 [−5.4, +8.6]	+2.0 [−5.3, +9.2]	+2.4 [−4.9, +9.7]	−3.9 [−10.3, +2.5]	+0.0 [−7.8, +7.8]
WMT25 ESA, primary metrics	−0.1 [−1.6, +1.5]	−0.4 [−2.1, +1.4]	+0.0 [−1.8, +1.8]	+0.2 [−1.8, +2.2]	−1.5 [−3.7, +0.7]
WMT25 MQM, LLM judges	+4.1 [−0.6, +8.8]	+6.4 [+1.0, +11.7]	+8.3 [+2.7, +14.0]	+7.0 [+1.0, +13.1]	+4.3 [−1.3, +10.0]
WMT25 ESA, LLM judges	−0.8 [−2.1, +0.4]	−0.1 [−1.4, +1.3]	+2.3 [+1.0, +3.7]	+5.6 [+3.5, +7.7]	+0.5 [−1.1, +2.1]
PPE, reward models	+1.6 [+0.7, +2.5]	+1.5 [+0.5, +2.5]	+1.5 [+0.5, +2.5]	–	−0.1 [−1.2, +1.0]
PPE, seven LLM judges	+0.0 [−0.3, +0.4]	−0.1 [−0.6, +0.3]	+0.4 [−0.1, +0.9]	+0.3 [−0.1, +0.7]	+0.5 [−0.1, +1.1]
WMT25 MQM, primary metrics, every system	+3.8 [−2.3, +9.9]	+7.0 [+0.9, +13.1]	+7.6 [+1.6, +13.6]	−0.3 [−6.0, +5.5]	+0.4 [−6.5, +7.4]
WMT25 ESA, primary metrics, every system	+0.8 [−1.0, +2.7]	+2.8 [+0.5, +5.1]	+2.9 [+0.6, +5.2]	+2.0 [−0.5, +4.5]	+1.3 [−1.2, +3.8]
WMT25 MQM, LLM judges, every system	+7.9 [+2.4, +13.5]	+8.3 [+2.3, +14.4]	+11.2 [+5.0, +17.5]	+8.6 [+2.2, +15.0]	+3.8 [−3.5, +11.0]
WMT25 ESA, LLM judges, every system	−0.2 [−1.0, +0.6]	+2.2 [+1.0, +3.3]	+4.7 [+3.6, +5.8]	+6.4 [+4.9, +7.9]	+1.5 [+0.1, +2.9]
half the clusters (PPE: 48 prompts)					
exploratory	+1.2 [−0.5, +2.9]	+1.6 [−0.2, +3.5]	+2.1 [+0.3, +4.0]	+1.0 [−0.6, +2.7]	+0.2 [−2.0, +2.3]
WMT23 MQM	+0.7 [−1.4, +2.8]	−0.6 [−3.6, +2.4]	−0.4 [−3.3, +2.5]	+1.0 [−0.9, +3.0]	+1.8 [−2.1, +5.6]
RewardBench 2	+1.3 [−0.1, +2.6]	+1.8 [+0.2, +3.5]	+1.8 [+0.2, +3.4]	+1.2 [−0.6, +3.0]	+1.2 [−0.6, +3.0]
PRM800K	+0.0 [−0.1, +0.1]	−0.2 [−2.9, +2.5]	−0.3 [−3.0, +2.4]	–	+0.0 [−0.3, +0.3]
WMT24 MQM	−2.7 [−6.2, +0.8]	−2.8 [−9.2, +3.7]	−1.3 [−7.8, +5.1]	−3.5 [−9.0, +2.0]	−6.0 [−13.0, +1.1]
WMT24 ESA	+2.0 [−0.3, +4.2]	+2.0 [−0.3, +4.3]	+2.1 [−0.3, +4.6]	+1.9 [−0.4, +4.2]	+0.2 [−2.6, +3.0]
WMT25 MQM, primary metrics	+0.3 [−8.2, +8.7]	+4.7 [−4.0, +13.4]	+5.2 [−2.8, +13.2]	−5.0 [−13.5, +3.5]	+4.0 [−7.5, +15.5]
WMT25 ESA, primary metrics	+0.9 [−1.2, +3.0]	+1.5 [−0.8, +3.9]	+1.8 [−0.6, +4.2]	+1.5 [−1.0, +4.1]	−0.4 [−3.9, +3.0]
WMT25 MQM, LLM judges	−0.7 [−7.2, +5.8]	+0.5 [−8.1, +9.1]	+1.5 [−6.4, +9.5]	+0.3 [−7.8, +8.5]	−4.5 [−13.3, +4.2]
WMT25 ESA, LLM judges	+1.1 [−0.3, +2.5]	+1.1 [−0.5, +2.7]	+1.6 [−0.1, +3.2]	+1.9 [−0.5, +4.2]	+1.1 [−1.0, +3.2]
PPE, reward models	+0.9 [−0.7, +2.5]	+0.1 [−1.6, +1.8]	+0.1 [−1.6, +1.8]	–	−0.9 [−2.8, +0.9]
PPE, seven LLM judges	+0.0 [−0.7, +0.8]	−0.2 [−1.2, +0.9]	+0.3 [−0.6, +1.3]	+0.1 [−0.8, +1.0]	+0.3 [−0.8, +1.3]
WMT25 MQM, primary metrics, every system	+0.5 [−9.8, +10.9]	+5.1 [−4.6, +14.7]	+4.8 [−4.6, +14.1]	−8.3 [−16.8, +0.1]	+0.4 [−12.3, +13.1]
WMT25 ESA, primary metrics, every system	−0.1 [−2.2, +2.1]	+3.1 [+0.8, +5.5]	+3.2 [+0.8, +5.5]	+1.2 [−1.5, +4.0]	+2.8 [−0.5, +6.0]
WMT25 MQM, LLM judges, every system	+1.2 [−4.3, +6.6]	−1.2 [−7.7, +5.3]	+0.5 [−5.8, +6.7]	−2.1 [−9.1, +4.8]	−1.3 [−8.1, +5.5]
WMT25 ESA, LLM judges, every system	+0.3 [−0.5, +1.0]	+3.4 [+2.0, +4.8]	+3.9 [+2.7, +5.2]	+3.6 [+2.2, +5.1]	+2.9 [+1.2, +4.7]

R EVERY GRADE OF THE FROZEN PROTOCOLS

Table 42 prints every grade the five protocols and the fifth's two addenda gave, in the order they were written. Each protocol was written after the grades of the protocols before it had been read and fixed its tests before the quantities they grade were computed; what each had read before it was frozen is stated in the table's caption and in each protocol's appendix. They test readings forward one protocol after another and are not one confirmatory family: the grades carry no correction for multiplicity, a grade of consistent leaves the reading open, and 12 of the 61 supported grades could not have come out not supported or contradicted (†). Table 1 gives the best-of-n estimates behind the first, third and fourth protocols' ordering and utility grades at half the clusters, the first protocol's with the recomputed intervals.

Table 42: **Every grade of every frozen test, in the order the protocols were written.** S supported, the estimate on the predicted side and its interval excluding the threshold; C consistent, on the predicted side, the interval covering it; N not supported, on the other side, the interval covering it; X contradicted, the interval excluding it on the other side. bon, best of n; pair, the pairwise choice; grp, the grouped mean. Each protocol was hashed before any quantity it grades was computed: protocols 1 and 4 before their data were opened, protocol 2 on sets already read with every candidate kept and its addendum after the first run's grades, protocol 3 after WMT24's MQM label files had been opened to count systems, protocol 5 before any judge had rated a response of the sets it grades, its addendum 5a after protocol 5's grades, before the judges' mean had been computed on WMT25, and its second addendum 5b after 5a's grades, before any jury had been computed (Appendices S, N, T, U and V); protocols 2 to 5b also carry a third party's time stamp, whose times are printed. Protocol 1 as its frozen run graded it; *supported under the later 300-split recomputation, the one grade it changes. †Could not come out not supported or contradicted: pooling's own regret at half lay below the threshold the gain is held to (M2's line, M3's 3 or 1 points, L4's 3 points), and no rule's gain over pooling can exceed that regret. ‡Graded unpaired on the MQM sets (Deviation 2). WMT25 leaves out English–Maasai (Deviation 1). The eight frozen sets of Figure 1b: WMT23, RewardBench 2, and WMT24's and WMT25's MQM and ESA sets, WMT25's with each panel. Estimates: Tables 44, 24, 30, 48, 53, 57, 60 and 61.

1. Validation, hashed 2026-09-15 (machine clock); half the clusters, frozen run's intervals

	WMT23		*RewardBench 2*	
	bon	pair	bon	pair
H1 grp's pick gains over r_T's	N	N	S	S
H2 and over the own statistic's	N	N	C	C
H3 grp orders evaluator pairs better	S	S	S	S
H4 at 8 clusters best of n picks worse than grp	S		C *	
H4 and has the largest error of the rules		S		N
H5 grp wins when the two rules disagree	N		S	
H6 pooling's limit error positive everywhere		S		S

2. Candidates, token 2026-09-18 22:21:58, and its addendum; 8 labelled clusters, best of n

	PPE	*exploratory*	*RB2*	*PRM800K*
grp's gain falls as candidates are removed	C	S	S	C
and is negative at four candidates	X			
r_W above grp with the fewest candidates	C	S	S	S
r_W above pooling at four candidates	S			

3. WMT24, token 2026-09-19 12:16:40; half the segments ($^+$ with the sentinel metrics)

	MQM		*ESA*		*MQM*$^+$		*ESA*$^+$	
	bon	pair	bon	pair	bon	pair	bon	pair
M1 advantage at least 8 points	C	S	S	S	S	S	C	S
M2 gain below the line fitted outside MT	S †	S †	S †	S †	S †	S †	S	S †
M3 gain below 3 (bon) or 1 (pair)	C	S †	C	C	C	S †	C	C
M4 gain falls from 32 segments to half	C		N		C		N	
M4 and the advantage rises	S		S		S		S	
label by the rule (M1 and M2 under best of n)	as the grades stand		confirmed		confirmed		as the grades stand	

4. WMT25, token 2026-09-25 06:35:45; half the segments; the year's primary metrics (left) and thirteen LLM judges (right)

	MQM		*ESA*		*MQM*		*ESA*	
	bon	pair	bon	pair	bon	pair	bon	pair
N1 advantage at half above zero	S	S	S	S	S	S	S	S
M1 advantage at least 8 points	C	N	C	C	N	N	X	X
M2 gain below the line fitted outside MT	C	C	S	S	C	C	C	S †
M3 gain below 3 (bon) or 1 (pair)	N	C	C	S	C	C	S	S †
M4 gain falls from 32 segments to half‡	N		N		C		N	
M4 and the advantage rises‡	C		S		N		N	
N4 best of n on 8 labelled segments below r_T	C		C		N		N	
N2a at 8 segments, $r(\mathrm{adv}, \mathrm{gain}) > 0$, all 15	S (pair S)				C (pair C)			
N2b r at 8 minus r at half > 0, all 15	C (pair C)				C (pair C)			
N5 ESA's gain at half above MQM's	N (pair C)				C (pair N)			
labels: ordering (N1); the gain tracks the advantage (N2)	confirmed; as the grades stand				confirmed; as the grades stand			
label of the third protocol's rule	as the grades stand		as the grades stand		failed		failed	

5. Best of 16 on PPE, token 2026-09-25 15:11:06; 8 labelled prompts unless stated

	seven LLM judges	*with the reward models*
L1 at half grp orders judge pairs better	C	S
L2 acc$^*_{eq}$'s pick beats r_T's	C	C
L3 and beats grp's	C	
L4 at half grp's gain below 3 points	S †	
L5 best of 16 on the labelled prompts picks worse than r_T	N	
label of the advice (L2)	as the grades stand	

5a. Its addendum on WMT25, token 2026-09-25 20:23:21; the thirteen LLM judges, best of n

	MQM	*ESA*
E1 the judges' mean picks better than r_T at 8 segments	S	X
E2 and better than the best judge chosen on all labels	C	X

5b. Its second addendum, token 2026-09-25 20:59:12; a jury of the three judges r_T ranks highest on 8 labelled units

	PPE, best of 16	*WMT25 MQM*	*WMT25 ESA*
J1 the jury picks better than r_T's single pick	S	S	S
J2 and better than the mean of all the judges	X	N	S

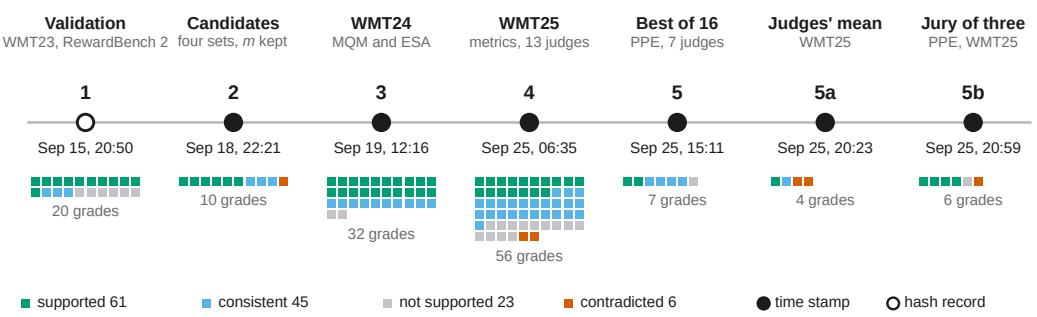


Figure 11: **The frozen protocols in time.** Each node is a protocol or addendum, in the order they were frozen, with the time of its record, UTC: protocol 1 a hash record by the machine clock, the others a third party's RFC 3161 time stamp. Below each, one square for each grade Table 42 prints for it, coloured by grade. Evenly spaced, not to scale in time (`scripts/figures/fig_timeline.py`, which reads the counts and times from the table).

Table 43: The hypotheses of the validation protocol, frozen before the validation data were read (Section 5), graded by the frozen code; configurations in parentheses. Differences in points, the estimate ± 1.96 corpus-bootstrap standard deviations of the mean over configurations. supp.: the sign predicted and the interval clear of zero; cons.: the sign predicted; not supp.: the other sign; contr.: the other sign and the interval clear of zero. H5: the share of disagreeing draws the grouped mean's evaluator wins and the configurations above one half, no interval; H6: whether pooling computed on every cluster orders some evaluator pair against the corpus's own pairwise payoff, a sanity check that fails where it orders every decisive pair as the payoff does. The protocol named no primary hypothesis and applied no correction, so the family of about two dozen readings is descriptive. Intervals from 200 corpus bootstraps of 300 splits each, the frozen run's 40 raised after the calibration (a deviation: Table 44). †The secondary panel the protocol fixed, RewardBench 2 without the authors' own directory, reported for comparison and not graded.

	Quantity	Exploratory (19)	WMT23 MQM (3)	RewardBench 2 (5)	RB2 w/o allenai† (5)
H1	grp−r_T, utility, best of n, half	+1.6 [−0.2, +3.5] cons.	−0.6 [−3.6, +2.4] not supp.	+1.8 [+0.2, +3.5] supp.	+0.6 [−0.7, +2.0] cons.
	grp−r_T, utility, pairwise choice, half	+0.9 [+0.4, +1.4] supp.	−0.3 [−1.0, +0.4] not supp.	+0.9 [+0.1, +1.7] supp.	+0.4 [−0.2, +1.0] cons.
H2	grp−own, utility, best of n, half	+0.6 [−0.3, +1.6] cons.	−1.6 [−4.5, +1.2] not supp.	+0.7 [−0.3, +1.6] cons.	+0.4 [−0.4, +1.1] cons.
	grp−own, utility, pairwise choice, half	+0.3 [−0.0, +0.5] cons.	−0.5 [−1.1, +0.1] not supp.	+0.4 [−0.1, +0.8] cons.	+0.2 [−0.1, +0.6] cons.
H3	grp−r_T, error, best of n, half	−5.0 [−6.6, −3.4] supp.	−12.4 [−14.6, −10.3] supp.	−1.8 [−2.7, −0.9] supp.	+0.3 [−0.4, +1.1] not supp.
	grp−r_T, error, pairwise choice, half	−7.9 [−9.4, −6.5] supp.	−14.6 [−16.2, −13.0] supp.	−1.4 [−2.4, −0.5] supp.	+0.3 [−0.5, +1.0] not supp.
H4	largest error at 8 clusters, pairwise	bon supp.	bon supp.	own not supp.	own not supp.
	bon−grp, utility, best of n, 8 clusters	−5.6 [−6.9, −4.3] supp.	−3.2 [−4.4, −2.0] supp.	−1.0 [−1.8, −0.3] supp.	−0.9 [−1.9, +0.0] cons.
H5	grp wins when the two disagree, half	63%, 13 of 17 supp.	49%, 1 of 3 not supp.	79%, 3 of 4 supp.	68%, 2 of 3 supp.
H6	r_T limit error > 0, pairwise	19 of 19 supp.	3 of 3 supp.	5 of 5 supp.	5 of 5 supp.

S THE VALIDATION SETS AND THEIR ADAPTERS

What was frozen, and when. Figure 11 places the seven frozen records in time. `prereg/PROTOCOL_validation.md` states the data, the rules, the budgets, the hypotheses and the grading of Section 5; `prereg/freeze.py` recorded its SHA-256 and those of the 13 files the analysis runs, the engine, the own statistics, the separation and the grading scripts, the WMT23 normaliser and the RewardBench 2 fetcher, at 2026-09-15 20:50:36 UTC, when neither validation directory existed. The normaliser and the fetcher had been written before the freeze and not run; the RewardBench 2 parser and the loader were written afterwards, since the format of the score files could not be seen without opening them. No frozen file changed afterwards. One parameter of the interval did, as a deviation: the frozen run averaged 40 half-splits in every resampled corpus where the estimate averages 300, and the calibration of Appendix AA found the resulting interval of the primary mean 1.47 times too wide at eight clusters and up to 2.6 times per configuration, the split-to-split noise of the naming entering every replicate; `scripts/decision/boot300.py` recomputes the replicates with 300 splits each by consuming the frozen bootstrap streams in their order, so the first 40 splits of every replicate are the frozen run's own (checked to the last digit on every primary configuration), the draws, the point estimates and the grading code being untouched. Table 43 carries the recomputed intervals; Table 44 carries the frozen ones, and the two grade 1 of the

Table 44: The hypotheses of the validation protocol, frozen before the validation data were read (Section 5), graded by the frozen code; configurations in parentheses. Differences in points, the estimate ± 1.96 corpus-bootstrap standard deviations of the mean over configurations. supp.: the sign predicted and the interval clear of zero; cons.: the sign predicted; not supp.: the other sign; contr.: the other sign and the interval clear of zero. H5: the share of disagreeing draws the grouped mean's evaluator wins and the configurations above one half, no interval; H6: whether pooling computed on every cluster orders some evaluator pair against the corpus's own pairwise payoff, a sanity check that fails where it orders every decisive pair as the payoff does. The protocol named no primary hypothesis and applied no correction, so the family of about two dozen readings is descriptive. Estimates over 300 half-splits; intervals from 200 corpus bootstraps of 40 splits each, the frozen run's own; Table 43 carries the intervals recomputed with 300. †The secondary panel the protocol fixed, RewardBench 2 without the authors' own directory, reported for comparison and not graded.

	Quantity	Exploratory (19)	WMT23 MQM (3)	RewardBench 2 (5)	RB2 w/o allenai† (5)
H1	grp$-r_T$, utility, best of n, half	+1.6 [−0.3, +3.5] cons.	−0.6 [−3.9, +2.7] not supp.	+1.8 [+0.1, +3.5] supp.	+0.6 [−0.8, +2.1] cons.
	grp$-r_T$, utility, pairwise choice, half	+0.9 [+0.4, +1.5] supp.	−0.3 [−1.0, +0.4] not supp.	+0.9 [+0.1, +1.7] supp.	+0.4 [−0.2, +1.1] cons.
H2	grp−own, utility, best of n, half	+0.6 [−0.4, +1.7] cons.	−1.6 [−4.7, +1.4] not supp.	+0.7 [−0.6, +1.9] cons.	+0.4 [−0.6, +1.3] cons.
	grp−own, utility, pairwise choice, half	+0.3 [−0.0, +0.5] cons.	−0.5 [−1.2, +0.1] not supp.	+0.4 [−0.3, +1.0] cons.	+0.2 [−0.2, +0.7] cons.
H3	grp$-r_T$, error, best of n, half	−5.0 [−6.7, −3.3] supp.	−12.4 [−14.7, −10.2] supp.	−1.8 [−2.7, −0.9] supp.	+0.3 [−0.4, +1.1] not supp.
	grp$-r_T$, error, pairwise choice, half	−7.9 [−9.5, −6.4] supp.	−14.6 [−16.4, −12.9] supp.	−1.4 [−2.4, −0.5] supp.	+0.3 [−0.5, +1.1] not supp.
H4	largest error at 8 clusters, pairwise	bon supp.	bon supp.	own not supp.	own not supp.
	bon−grp, utility, best of n, 8 clusters	−5.6 [−7.3, −3.8] supp.	−3.2 [−5.8, −0.6] supp.	−1.0 [−2.7, +0.6] cons.	−0.9 [−2.7, +0.8] cons.
H5	grp wins when the two disagree, half	63%, 13 of 17 supp.	49%, 1 of 3 not supp.	79%, 3 of 4 supp.	68%, 2 of 3 supp.
H6	r_T limit error > 0, pairwise	19 of 19 supp.	3 of 3 supp.	5 of 5 supp.	5 of 5 supp.

21 interval-read cells of the three graded columns differently (H4's utility contrast on RewardBench 2, consistent to supported), and none of the seven labels of the ungraded panel. The exploratory column of either table is the same code graded on the primary configurations after the freeze, a reference and not evidence.

WMT23 MQM. From the mt-metrics-eval archive already on disk for WMT22: every system rated by MQM that is neither a human reference nor marked as an outlier, on the source segments every such system is rated on, en-de 460 segments of 12 systems, he-en 820 segments of 12 systems, zh-en 1177 segments of 15 systems, 461 documents in all; the label is the official MQM segment score, and the mean item label of every system equals the archive's MQM system score to 10^{-14}. The panel is every metric whose basename is in the archive's `WMT23_PRIMARIES`, scored with the standard reference (refA for en-de and zh-en, refB for he-en) or with the source, and every such metric scores every item, 32 a language pair, among them the random baseline the task lists as primary. The own statistic, acc^*_{eq} inside the source segment, is checked against the item-level implementation to 10^{-15} on random masks.

RewardBench 2. The benchmark and the per-completion scores of `allenai/reward-bench-2-results` at the revisions in the fetch record: 1763 prompts in five of its six subsets, Ties excluded by the protocol since its prompts carry several correct completions, 160 to 495 a subset, each with one chosen and three rejected completions. The 14 API judges record no per-completion score and are excluded by the panel rule, which leaves 174 sequence classifiers, 132 of them in the benchmark authors' own directory. Two properties of the release were found by the adapter and are recorded rather than repaired by guesswork. Prompt identifiers repeat across subsets, and 179 of 188 score files label their first ten Factuality entries with the identifiers of entries 30 to 39; every file is therefore aligned with the benchmark by position, and every entry of every file is checked against the benchmark's row on its subset, its number of correct completions and its prompt text, with no mismatch. And the release's own per-prompt outcome differs from the strict one of the protocol only on prompts where a rejected completion ties the chosen one: it counts 295 such model and prompt pairs as correct, what taking the first of the tied completions gives when the chosen one is listed first, and 26 as one half, which moves no model's accuracy on a subset by more than 2.5 points. The protocol's own statistic, the strict accuracy, is the one scored.

Families. 132 of the 174 models are the benchmark authors' training runs, and a panel that holds many checkpoints of one pipeline hands every rule pairs that differ in little but a seed or a learning rate, which is where a grouped statistic that reads within-prompt differences may gain or lose most. Beside the panel without the authors' directory, which the protocol fixed in advance, the frozen analysis was

Table 45: RewardBench 2 under panels balanced over model families, at half the prompts and at eight, 300 splits, the frozen analysis unchanged. A family is the organisation in the model's name, the benchmark authors' directory split into its three naming series; one member of every family, or at most two, is drawn at random, and the analysis is run on every draw. Err: the share of evaluator pairs on which pooling / the grouped mean selects the evaluator that is worse on the evaluation half, under best of n; grp$-r_T$ err: the paired difference, with the interval of Table 43 for a single panel and the range over draws otherwise; grp$-r_T$: the paired utility difference of the selected evaluators, best of 4 recovery and pairwise accuracy, in points; pos. / clear: the draws on which the best-of-4 difference is positive and on which its interval is clear of zero; $r_W - r_T$: the within-cluster coefficient's evaluator against pooling's; pos. / above / below: the draws on which the best-of-4 difference is positive, on which its interval lies wholly above zero and wholly below zero. Run after the grading, not part of the frozen protocol.

Panel	b	Ev.	draws	err r_T / grp	grp$-r_T$ err	grp$-r_T$, best of 4	pos. / above / below	grp$-r_T$, pairwise	$r_W - r_T$, best of 4
every model	half	174	1	17.7 / 15.9	−1.8 [−2.7,−0.9]	+1.8 [+0.2,+3.5]	–	+0.9 [+0.1,+1.7]	+1.3
every model	8	174	1	26.2 / 30.0	+3.8 [+3.3,+4.3]	−2.4 [−3.8,−1.0]	–	−0.8 [−1.6,−0.1]	+0.9
w/o authors' models	half	42	1	10.6 / 10.9	+0.3 [−0.4,+1.1]	+0.6 [−0.7,+2.0]	–	+0.4 [−0.2,+1.0]	+0.7
w/o authors' models	8	42	1	19.3 / 23.5	+4.1 [+3.6,+4.7]	−1.7 [−3.2,−0.2]	–	−0.5 [−1.3,+0.2]	+0.9
one per family	half	21	20	12.8 / 13.1	+0.3 (−0.7 to +1.5)	+0.4 (−0.4 to +1.7)	12 / 0 / 0	+0.1 (−0.4 to +0.7)	+0.6 (−1.2 to +1.5)
one per family	8	21	20	22.1 / 26.7	+4.6 (+3.7 to +5.2)	−3.0 (−3.5 to −2.5)	0 / 0 / 20	−1.2 (−1.5 to −1.0)	−0.3 (−0.7 to +1.0)
two per family	half	32	10	13.1 / 13.1	+0.0 (−0.4 to +0.5)	+0.4 (−1.0 to +1.0)	8 / 0 / 0	+0.2 (−0.6 to +0.5)	+1.0 (−0.7 to +1.7)
two per family	8	32	10	22.1 / 26.1	+4.1 (+3.4 to +4.6)	−2.6 (−2.9 to −2.2)	0 / 0 / 10	−1.0 (−1.2 to −0.8)	+0.1 (−0.6 to +1.6)

run after the grading on panels balanced over families, a family being the organisation in a model's name with the authors' directory split into its three naming series, 21 families in all: one member of each drawn at random, 20 draws, and at most two, 10 draws (Table 45; `validation_sets.py`, `rb2fam1`, `rb2fam2`). With one model per family the grouped mean's evaluator earns +0.4 points of best-of-4 recovery over pooling's at half the prompts, from −0.4 to +1.7 over the draws, positive on 12 and with an interval clear of zero on 0, and +0.1 of pairwise accuracy; in selection error the grouped mean selects the worse evaluator on 13.1% of pairs against 12.8% for pooling, a gap of +0.3 points where the full panel's is −1.8. With at most two per family the utility gain is +0.4 (−1.0 to +1.0) and the error gap +0.0. The competitive and cross-family readings of the primary panels (Appendix I) stratify pairs; these panels remove the families' multiplicity before any pair is formed, and they are the reading under which the RewardBench 2 utility result of Table 43 is to be bounded.

T The Third Protocol: Ordering and Choice on an Unread Translation Year

Why, and what was frozen. The main line of Section 7 was read after the first two protocols were graded, and the third reading of the draft asked, from four sides, for one prediction of it fixed before the data were opened. The clause specific enough to fail is the one on machine translation. `prereg/PROTOCOL_mt_parting.md` states it for the WMT24 shared task; the protocol, an adapter, a runner and a grader were hashed together with every module they import (`prereg/FREEZE_mt_parting.txt`, 2026-09-19 12:17:03 UTC by the machine's clock), and an RFC 3161 token of the record was issued by a third party (12:16:40 UTC by its clock; the two clocks differ by less than half a minute). The analysis itself is the one frozen for the first protocol: `targets.py` and its modules have the hashes of that record, and the runner registers the new set at run time without changing a frozen file. The adapter ran at 12:17:11 and the analysis from 12:17:52.

What of WMT24 had been opened. Not nothing, and the protocol says so. On 2026-09-13 a script that was acquiring label data for the census of Table 63 read the MQM segment scores and ratings of en-de, en-es and ja-zh and printed the number of source segments, the rated systems with their numbers of rated segments, a check that the ratings sum to the scores, the raters' segment counts and the number of metric files; the acquisition stopped there, and WMT24 entered no table, analysis or number of the paper. On the day of the freeze the file names of two directories and mt-metrics-eval's metadata file, which names the year's primary metrics, standard references and outlier system, were read. No metric score of the year and no ESA score had been read, and no statistic relating an evaluator to a label had been computed on a WMT24 item. The project's command history, that of its subagents included, was searched for the year before this was written.

Data and panels. The WMT24 metrics task (Freitag et al., 2024) as the mt-metrics-eval v2 archive holds it, the eleven language pairs with a segment-level human score: MQM for en-de, en-es and ja-zh, the primary set, and error span annotation (ESA; Kocmi et al., 2024) for the other eight, a secondary set graded by the same rules because the clause was read on MQM labels and whether it carries to another annotation protocol was not known. Configurations are built exactly as for WMT23 (run on the WMT23 files, the adapter reproduces the WMT23 tables to the last digit): the rated translations of the systems that are neither human references nor the outlier system, on the source segments rated for every kept system; cluster the source segment, block its document, label the archive's segment score, own statistic acc^*_{eq}. A shortfall rule for language pairs with fewer than 300 such segments was fixed in advance and did not act. The panel is every metric among the year's primaries, under the standard reference or the source, that scores every item. That list holds the three sentinel metrics of Perrella et al. (2024b), which the task included as baselines to probe its meta-evaluation; two of them are constant inside a source segment by construction, so a within-segment rank or Pearson coefficient gives them nothing, a pooled coefficient credits their between-segment signal, and the task's own tie-calibrated acc^*_{eq} credits them the share of within-segment ties (the protocol's text, which calls them the organisers' and says no statistic grouped on the segment can credit them, is corrected here; the correction changes no graded quantity). They are left out of the graded panel, so that no ordering advantage below is owed to them, and the panels that keep them are graded beside it (Table 49). Table 47 lists the configurations.

Hypotheses and grading. In points, configurations weighing equally, the advantage is pooling's selection error over evaluator pairs less the grouped Spearman mean's and the gain is the paired utility difference of the evaluators the two rules select. M1: at half the segments the advantage is at least 8 points. M2: at half the gain is below the line of gain on advantage fitted at half on the 21 earlier configurations outside machine translation, taken as fixed ($+0.33 + 0.56\times$ advantage under best of n, $+0.69 + 0.06\times$ advantage under the pairwise choice; residual standard deviations 5.1 and 1.6 points, which the interval of the gain does not carry). M3: at half the gain is below 3 points of best-of-n recovery and 1 point of pairwise accuracy. M4: under best of n the gain at 32 segments exceeds the gain at half while the advantage at half exceeds the advantage at 32. An estimate is supported when it lies on the predicted side of its threshold and its interval, the estimate ± 1.96 standard deviations of the corpus bootstrap of the set mean (200 corpora of 40 half-splits, at 8, 32 and half), excludes the threshold; consistent when the interval covers it; not supported when the estimate lies on the other side; contradicted when the interval excludes the threshold on that side. The protocol calls the parting confirmed on a set when M1 and M2 are both supported under best of n and failed when either is not supported or contradicted, and commits to withdrawing the clause if the primary set fails. It also states what three configurations can show: on WMT23 the standard deviation of the set-mean gain at half was 1.5 points under best of n, so M3's margin there can be supported only by an estimate near or below zero and M4's gain contrast, +2.0 against -0.1 on the six earlier configurations, was more likely to come out consistent than supported.

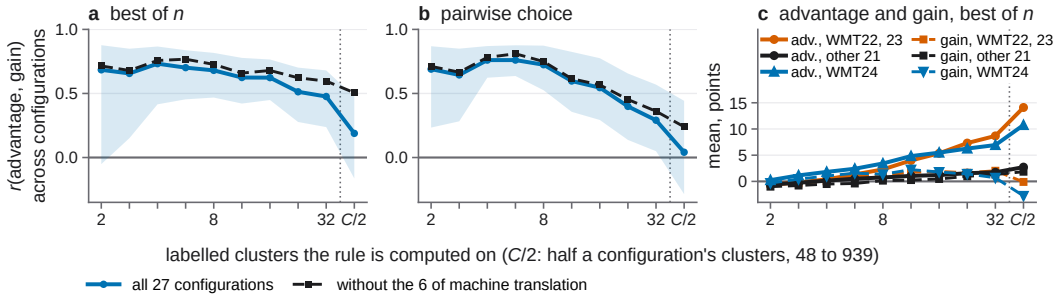


Figure 12: **Ordering and choice by label budget, read after the fact.** Left and middle: across configurations, the correlation between the grouped mean's ordering advantage over pooling and the utility gain of the evaluator it selects, by the clusters the rule reads, on all 27 configurations and without the 6 of machine translation ($C/2$: half a configuration's clusters; bands: 2.5th to 97.5th percentiles over resampled configurations, descriptive). Right: mean advantage and mean gain under best of n for those 6, for the other 21 and for the three WMT24 MQM language pairs of the frozen test.

Table 46: The third protocol's grades, every set: WMT24's three MQM language pairs (primary) and eight ESA pairs, each with the panel that leaves out the year's three sentinel metrics (graded estimates) and the one that keeps them (grades). Set means at half the segments in points: the grouped mean's ordering advantage over pooling and the utility gain of its evaluator, ± 1.96 standard deviations of the frozen corpus bootstrap (200 corpora of 40 splits); M2's line value in parentheses. supp. supported, cons. consistent, not supp. not supported. The headroom, pooling's regret on the evaluation half (read after the grading), bounds the gain: under both payoffs it lies below M2's line on both sets, so M2 could not come out not supported and the label could fail only through M1, and under the pairwise choice it also lies below M3's margin on the MQM set; without HANNA relevance the best-of-n line gives +3.5 and +3.7 at the two sets' advantages, where M2 would be consistent. Full table: Table 48.

	MQM (3) estimate	grade	ESA (8) estimate	grade	MQM +sentinels	ESA +sentinels
best of n						
M1 advantage ≥ 8	+10.8 [+6.7, +14.8]	cons.	+11.3 [+9.5, +13.2]	supp.	supp.	cons.
M2 gain below the line	−2.8 [−9.2, +3.7] (+6.3)	supp.	+2.0 [−0.3, +4.3] (+6.6)	supp.	supp.	supp.
M3 gain below 3	−2.8 [−9.2, +3.7]	cons.	+2.0 [−0.3, +4.3]	cons.	cons.	cons.
M4 gain falls, 32 to half	+3.5 [−2.5, +9.4]	cons.	−0.4 [−2.4, +1.6]	not supp.	cons.	not supp.
M4 advantage rises	+3.8 [+2.0, +5.6]	supp.	+5.8 [+4.5, +7.1]	supp.	supp.	supp.
headroom at half, not graded	5.6		6.5			
pairwise choice						
M1 advantage ≥ 8	+15.9 [+11.9, +19.9]	supp.	+12.4 [+10.1, +14.8]	supp.	supp.	supp.
M2 gain below the line	−0.2 [−1.1, +0.7] (+1.7)	supp.	+0.6 [+0.1, +1.1] (+1.5)	supp.	supp.	supp.
M3 gain below 1	−0.2 [−1.1, +0.7]	supp.	+0.6 [+0.1, +1.1]	cons.	supp.	cons.
headroom at half, not graded	0.9		1.0			
label by the rule	as the grades stand		confirmed		confirmed	as the grades stand

The primary set. Table 48; half of a configuration is 241 to 306 segments and the graded panels hold 23 metrics. Under best of n the advantage at half is +10.8 $[+6.7, +14.8]$ points, above its threshold with an interval that reaches below it (M1 consistent), and the gain is -2.8 $[-9.2, +3.7]$, below the line's +6.3 (M2 supported) and below the margin of 3 with an interval that reaches above it (M3 consistent). Under the pairwise choice all three are supported: +15.9 $[+11.9, +19.9]$ points of advantage and a gain of -0.2 $[-1.1, +0.7]$ against +1.7 on the line and a margin of 1. By the protocol's rule the parting stands as the grades stand on the primary set, and we claim the grades and no label. Its intervals are about twice as wide as WMT23's.

By language pair the advantage under best of n is +7.0 to +17.8 points and the gain -11.3 to +1.5 (Table 47). The set mean of the gain rests on ja-zh, where the metric the grouped mean selects is worth less than pooling's by more than ten points: leaving one language pair out, outside the protocol, the mean gain runs from -4.9 to +1.5 and the mean advantage from +7.2 to +12.6, and 3 of the 3 gains lie below the line at their own advantage. Along the budget the advantage rises from +3.4 points at eight segments to +6.9 at 32 and +10.8 at half while the gain goes from +1.4 to +0.7 and -2.8 (M4: the rise supported, +3.8 $[+2.0, +5.6]$; the fall consistent, +3.5 $[-2.5, +9.4]$), the shape of the six earlier configurations (Figure 12, right). Not graded: at half the pairwise accuracy on the labelled segments orders +12.0 points better than pooling with a gain of -1.3, the within-cluster coefficient +2.5 and -2.7, and the task's own statistic -2.0 and -3.5.

With the sentinels. With the three sentinel metrics in the panel the advantage is larger, +13.2 $[+9.2, +17.2]$ points under best of n (M1 supported) and +18.4 under the pairwise choice, and the gain is the same to the first decimal, -2.8 $[-8.7, +3.2]$ and -0.2; by the protocol's rule the parting is confirmed on that panel. It is the secondary reading the protocol made it: the ordering advantage the sentinels add is the published point of those metrics, and it adds nothing to the choice.

The ESA language pairs. Table 48, second column; half of a configuration is 283 to 974 segments and the graded panels hold 20 metrics. Under best of n the advantage at half is +11.3 $[+9.5, +13.2]$ points (M1 supported), by language pair +3.3 to +16.3, and the gain +2.0 $[-0.3, +4.3]$ against +6.6 on the line (M2 supported; M3 consistent), by language pair -0.9 to +5.0, 7 of the 8 below the line at their own advantage. Under the pairwise choice the advantage is +12.4 $[+10.1, +14.8]$ (supported) and the gain +0.6 $[+0.1, +1.1]$ against +1.5 (M2 supported; M3 consistent). From 32 segments to half the advantage changes by +5.8 $[+4.5, +7.1]$ (supported) and the gain falls by -0.4 $[-2.4, +1.6]$ (not supported). By the protocol's rule the parting is confirmed on this set; with the sentinels in

the panel the advantage is +8.9 $[+7.2, +10.6]$ and the gain +2.0 $[-0.2, +4.2]$ (as the grades stand). What the rule's label turns on is whether M1's interval clears its threshold under best of n, and that changes with the panel in opposite directions on the two sets: without the sentinels the label is as the grades stand on MQM and confirmed on ESA, with them confirmed on MQM and as the grades stand on ESA, where the advantage is smaller with them. The same in all four runs: the advantage is estimated above 8 points under both payoffs, it rises from 32 segments to half (supported every time), and the gain lies below the earlier line, M2 supported under both payoffs. What differs by annotation protocol: on the MQM language pairs the gain is negative in estimate and falls from 32 segments to half; on the ESA ones it is positive on 7 of 8 language pairs, clear of zero under the pairwise choice, and does not fall (-0.1, +1.6 and +2.0 at eight segments, 32 and half, the advantage +1.9, +5.5 and +11.3). So the registered reading, large ordering advantages that come with no gain, describes the MQM sets; on every set of the year the advantage comes with less gain than the same advantage does outside machine translation.

Which metrics the rules name. Table 50, read after the grading, gives the metric each rule names reading every segment, which the protocol's Section 7 promised. The MQM set's negative gain is ja-zh's, and there the grouped mean names GEMBA-ESA on 83% of the selection halves, a metric whose top score is tied among several of the segment's 13 translations on 93% of segments (7.3 on average when tied), which the per-segment Spearman mean over midranks credits and best of n pays as a pick at random inside the tied block; it ranks 7 of 23 by best-of-n recovery, while pooling names the best metric on 97% of halves. On two ESA language pairs it is pooling that names GEMBA-ESA.

Other panels. Table 51, not graded and read after the grading, reruns the frozen analysis on other panels. Without GEMBA-ESA the MQM set's advantage under best of n stays at +11.9 points and its gain moves to -0.3 $[-7.0, +6.4]$ (+17.3 and +0.4 under the pairwise choice): most of the negative gain was that metric's, and a large advantage with a gain near zero remains. On the same en-es systems and segments the MQM and the ESA label give gains of +5.5 and +5.8, so on those items the annotation protocol does not separate the two sets. With reference-free metrics only, the ESA set's gain is +4.0 $[+1.7, +6.3]$, clear of zero, and the MQM set's best-of-n advantage shrinks to +5.8.

What the headroom left M2 to test. Read after the grading. The gain of any rule over pooling cannot exceed the headroom, pooling's regret on the evaluation half, on any split, and on WMT24 the headroom at half was 5.6 points on the MQM set and 6.5 on the ESA set under best of n (0.9 and 1.0 under the pairwise choice), below the line's values at the observed advantages (Table 46). Given those headrooms no set mean of the gain could have come out above the line, and under the pairwise choice none on the MQM set above M3's margin either, so at the set level M2 graded chiefly that regret, which the protocol did not state as a prediction; M1 and M4 carried the rest. Under best of n the year's regret is at the level of the configurations outside translation (6.2 points), above WMT22 and WMT23's 4.0. The best-of-n line itself rests on HANNA relevance: refitted without it, its value is +3.5 at the MQM advantage and +3.7 at the ESA one (standard error 2.1 for the full line), where M2 would be consistent on both and the ESA label would stand as the grades stand; without any other configuration it stays above the MQM gain's interval, and under the pairwise choice it does so without any one configuration.

Deviations. None in what was run: no frozen file was changed after the freeze, each set was run once, and the grader refits the line from the earlier result files and stops if its constants differ from the frozen ones. Two statements of the protocol's text are corrected above (the sentinels' authors, and which grouped statistics credit them), and Table 47 prints the mean selection half where the adapter's card holds the smallest; neither changes a grade. The metrics each rule names, which the protocol's Section 7 promised with the grades, were computed only after them (Table 50). The leave-one-language-pair-out ranges above are outside the protocol and marked so.

U The Fourth Protocol: The Whole Line on WMT25, with Current LLM Judges

Why, and what was frozen. After the third protocol was graded, two further parts of the reading of Section 7 were taken on configurations the project had already read: that the gain follows the

ordering advantage with few labelled clusters and much less with many, and that best of n on few labelled clusters leaves the most regret. Simulated readers of an earlier draft (AI use statement) also found that no panel of current LLM judges scored against human labels had entered any analysis. The WMT25 metrics task (Lavie et al., 2025) answers both: it had entered no analysis, and beside the year's primary metrics, several of them built on LLMs (eleven in the findings paper, twelve in mt-metrics-eval, three names differing), it scores thirteen general-purpose LLMs that the organisers prompted as judges, zero-shot, with an ESA-like prompt that shows the source and not the reference. The protocol (`prereg/PROTOCOL_wmt25.md`), the two adapters, the runner and the graders were hashed with every module they import and with the archive the year was extracted from under an RFC 3161 token issued at 06:35:45 UTC (`prereg/FREEZE_wmt25.txt`; the record's own field, 2026-09-25 06:35:52, is the machine's clock, which ran 7 to 29 seconds ahead of the authority's throughout the project, so the token times are the ones to read).

What of WMT25 had been opened. Before the freeze: the names and sizes of the extracted year's files, never their content; mt-metrics-eval's metadata code; the two findings papers (Kocmi et al., 2025; Lavie et al., 2025), for how the scores were made; and, since the sixteenth reading's literature check, the metrics paper's published rankings by acc^*_{eq} inside the source segment. The protocol summarised those as placing the LLM judges in the bottom segment-level clusters, which is wrong for one of them: Gemini 2.5 Pro ranks first at the segment level, GPT-4.1 near the middle and most of the others low; the protocol also says that paper reports no pooled segment-level statistic, while it reports one on the general-task data, a best-threshold F1 for flagging catastrophic translations; these two are the statements of the protocol's text corrected here. That paper reports no pooled segment-level correlation on the general-task data, no selection from labelled segments and no reranking utility, so none of the quantities graded here, and every direction and threshold came from the project's earlier sets. The file system corroborates the record: the first read of any extracted file is the adapter's, 54 seconds after the freeze record's own time, and when this was written 1,764 of the 2,257 files had not been read at all.

Data and panels. The WMT25 general translation task (Kocmi et al., 2025) as the mt-metrics-eval v2 archive holds it: MQM for ja-zh and en-ko and ESA (Kocmi et al., 2024) for fourteen other language pairs, on segments that are usually paragraphs, with the first of the two ESA ratings, the score the archive names. The third protocol's adapter, unchanged, builds the items. Human evaluation covered the half of each language pair's inputs translated most diversely (pairwise chrF), and every system is rated on the same segments, so the adapter's shortfall rule, which drops the systems with the fewest rated segments one at a time and breaks ties by name until 300 segments or eight systems remain, gained no segment and kept on every language pair the eight systems last by name, 112 to 231 segments each. Among them are systems built on models that are also LLM judges, so that a judge scores its own translations on 7 language pairs: Qwen3-235B on 5, Llama-4-Maverick and Llama-3.1-8B on one each; with every system kept, 3 to 7 judges do on each pair, Claude 4, GPT-4.1 and Gemini 2.5 Pro on all fifteen. Gemini 2.5 Pro scores in its reasoning mode, which it uses by default. A rerun with every rated system is reported below. The ESA label is the first of the two ratings, as the archive names it; the two agree within a segment at 0.04 to 0.39 on 11 of the 13 language pairs (Pearson correlation of the ratings centred on their segment means, `results/wmt25_described_r17.json`), and the metrics paper's second version grades against their mean, which a second rerun uses. Two panels score them: the metrics mt-metrics-eval lists as the year's primaries under the standard reference or the source, 11 to 12 of which score every item, and the thirteen LLM judges, whose scores take few distinct values and tie at the top of a segment often (Tables 54 and 56). Deviation 1, recorded and time-stamped before any run: on English–Maasai no metric of either panel scores every item, so the language pair is left out and fifteen configurations a panel remain.

Hypotheses and grading. Advantage and gain as in the third protocol. N1: at half the segments the advantage is positive, on each set and under each payoff. N2a: over every configuration of a panel, the correlation of advantage and gain at eight segments is positive; N2b: it exceeds the same correlation at half; the interval of both is the 2.5th to 97.5th percentile of 10,000 hierarchical resamples, configurations drawn with replacement and one corpus-bootstrap replicate drawn for each. N4: at eight segments under best of n, best of n on the labelled segments picks a worse evaluator than pooling. N5: the gain at half is larger on ESA than on MQM. The third protocol's M1 to M4

are graded by its own grader with its frozen line, M4's interval under Deviation 2 on the MQM sets. Labels: the first clause is confirmed when N1 is supported on both sets under both payoffs, the second when N2a and N2b are both supported under best of n, and each fails when a grade that enters it is not supported or contradicted. The protocol stated beforehand that its grader, run on WMT24's eleven configurations, gave N2 as consistent, so that with sixteen language pairs, fifteen after Deviation 1, N2 was more likely to come out consistent than supported.

Results. Every grade is in Tables 42 and 53, and the configurations in Table 54. With both panels the ordering clause is confirmed: N1 is supported on both sets under both payoffs. With both panels whether the gain tracks the advantage (N2) stands as the grades stand. N2a is supported with the primaries, its lower bound +0.03, and consistent with the LLM judges; N2b is consistent with both. With the primaries the correlation at half (0.56) is as high as at eight (0.58), within the thirteen ESA language pairs it rises from eight segments to half in both panels (Figure 13), and N2b's point estimate turns negative without any one of five language pairs; with the LLM judges it falls from 0.44 to 0.04 over all fifteen and stays positive without any one. Fifteen configurations determine these correlations weakly: at eight segments a configuration's gain varies more across its own corpus bootstrap than across configurations with the primaries (standard deviations 3.8 and 2.8 points) and somewhat less with the LLM judges (2.7 and 3.3), and within a configuration the replicates of advantage and gain correlate at 0.44 and 0.34, so part of any correlation across configurations is shared noise; the ceiling a linear relation could show under that noise is 0.00 at eight segments and 0.31 at half with the primaries, 0.36 and 0.00 with the LLM judges, against 0.88 and 0.79 on the earlier sets (`results/wmt25_noise_ceiling_r17.json`); with the primaries the gain's reliability across configurations is nil, so N2a's supported 0.58 is what the shared noise alone can produce and is no evidence that the gain tracks the advantage. N2b is consistent in every run, the reruns included (+0.07 to +0.52 outside the frozen primaries). N4 is consistent with the primaries and not supported with the LLM judges, whose pooled pick at eight segments leaves the most regret of the six rules on MQM. N5 is not supported under best of n with the primaries: the MQM gain at half, +4.7, lies above the ESA gain and comes from en-ko alone. The third protocol's rule labels the primaries' sets as the grades stand (M1 consistent on both, M2 consistent on MQM and supported on ESA, M3 not supported on MQM) and the LLM judges' sets failed (M1 not supported on MQM and contradicted on ESA, their advantage at half, +6.4 and +3.5 points, lying below eight). M2 is at least consistent on all four sets, and unlike on WMT24, pooling's regret at half lay above the line's value on all four under best of n, so M2 could have failed there; under the pairwise choice it lay below on the LLM judges' ESA set (Table 42). The line owes its height to HANNA relevance: refitted without that configuration it gives 3.1 and 3.1 at the primaries' MQM and ESA advantages, where M2 would be not supported and consistent, and at WMT24's advantages, where both M2 grades would be consistent (`results/wmt25_described_r17.json`). Deviation 2: on en-ko two resampled corpora left fewer than 33 labelled segments, so that configuration holds 198 bootstrap replicates at 32 segments and 200 at half; the third protocol's grader, which pairs the two by position for M4, stopped, and M4 is graded with the two set means' bootstrap spreads combined as if independent, a change recorded and time-stamped before any grade of M1 to M4 was computed, but after the new hypotheses' grades and the runs' summary lines had been read, which between them carry the point estimates of M1 to M4; only M4's interval and grade were new, M4 enters no label, and the unpaired interval is the wider, so it can turn what a paired one would grade contradicted into not supported (`prereg/DEVIATIONS_wmt25.md`). Deviation 2 was appended to the file whose first 1556 bytes Deviation 1's record hashes; those bytes are kept as `prereg/DEVIATIONS_wmt25_d1.md`.

Described, not graded. Over the 8 set, panel and payoff cells of the frozen run, M3, the third protocol's test of a small gain, is supported on 3, consistent on 4 and not supported on 1. Within the thirteen ESA language pairs with the primaries the gain tracks the advantage at half ($r = 0.82$), so the registered reading that large ordering advantages come with no gain holds for the ESA set mean and not across its language pairs. Most of the LLM judges score in 8 to 26 distinct values, and for 12 of the thirteen several systems share the top score on most segments; Gemini 2.5 Pro, whose segment-level lead the metrics paper had published before the freeze, is the best judge on 9 of the 15 configurations. With eight labelled segments acc^*_{eq} picks a better judge than pooling by +7.0 $[+1.0, +13.1]$ points on MQM and +5.6 $[+3.5, +7.7]$ on ESA, naming the best judge on 62% of ESA draws against pooling's 34%, while the plain pairwise accuracy gains +2.3 and the grouped mean -0.1 on ESA; the fifth protocol graded this description on PPE (Appendix V). With every system

kept (below), the grouped mean's pick with eight labelled segments is worth about 28 of pooling's labelled segments on ESA and 53 on MQM.

Reruns outside the protocol. Read after the grading, and decided then (the AI use statement says by whom), the frozen analysis was rerun on tables that change one choice each (`scripts/normalize/wmt25_variants.py`, `scripts/decision/run_wmt25_variants.py`), graded by the same graders and reported whatever came out (Table 52). With every rated system kept, 18 to 20 a language pair on the same segments, the ordering advantage at half is +10.3 and +11.8 points with the primaries (supported and supported), the ESA gain is +3.1 $[+0.8, +5.5]$ against the line's +6.9, M1 and M2 both supported, so that the third protocol's rule would label the ESA set confirmed, the correlation of advantage and gain is 0.69 at eight segments and 0.57 at half, and best of n on eight labelled segments, now among eighteen candidates, is ahead of pooling on both sets (+0.4 and +1.3 points). With the merged ESA score as the label, the one the metrics paper's second version grades against, the ESA advantage is +10.8 points and the gain +1.0 $[-1.2, +3.1]$ against the line's +6.3, M1 and M2 both supported, so that the third protocol's rule would label the set confirmed; the correlation falls from 0.47 at eight segments to 0.21 at half. Without Gemini 2.5 Pro the LLM judges' ordering advantage at half is +7.8 and +3.9 points, and at eight segments the pairwise statistics still pick better than pooling: acc^*_{eq} by +6.5 and +4.6 points, the pairwise accuracy by +8.3 and +2.0, the grouped mean by +7.3 and +0.4. With every rated system and the LLM judges, the ordering advantage at half is +7.8 and +2.7 points (supported and supported), but the ESA gain at half is +3.4 $[+2.0, +4.8]$, above the line's +1.8 at that small advantage (M2 contradicted, the one M2 of any WMT25 run that is), and at eight segments every within-segment rule but r_W picks better than pooling on ESA: acc^*_{eq} by +6.4 $[+4.9, +7.9]$, the pairwise accuracy by +4.7, the grouped mean by +2.2 $[+1.0, +3.3]$; the correlation of advantage and gain is 0.21 at eight segments and -0.28 at half. On those pools 3 to 7 judges score their own translations on each pair; with the systems built on a judge's model left out (seven to fifteen a language pair, English–Maasai back in, since every judge then scores all of it), the ESA gain is +0.9 $[-0.6, +2.3]$ against the line's +2.8, M2 supported, and acc^*_{eq}'s pick at eight segments gains +8.9 and +6.4 points. With the merged ESA label and the LLM judges the ESA advantage is +4.8 points, M1 contradicted and M2 supported, and acc^*_{eq}'s pick at eight segments gains +5.5 $[+3.7, +7.4]$.

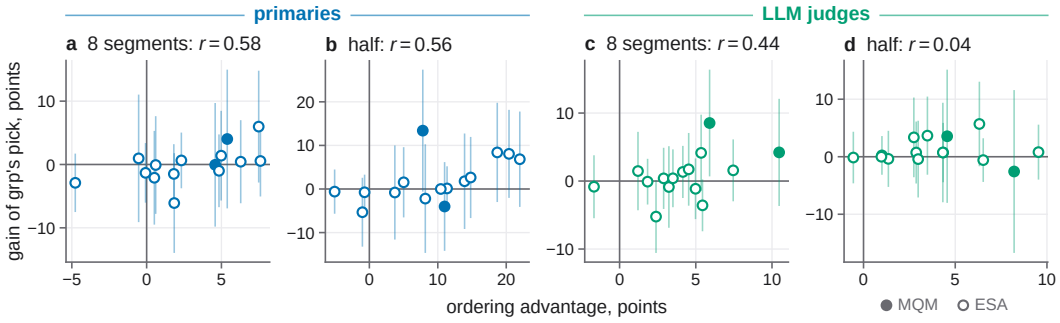


Figure 13: **The fourth protocol's configurations.** On WMT25, one point per language pair, the grouped mean's ordering advantage over pooling against the utility gain of the evaluator it selects, under best of n, at eight labelled segments and at half ($C/2$), for the year's primary metrics (blue) and for the thirteen LLM judges (green); filled, the two MQM language pairs; bars, 1.96 corpus-bootstrap standard deviations of the gain. The correlation over the fifteen configurations of a panel, in each title, is N2a at eight segments and enters N2b.

V The Fifth Protocol: Choosing an LLM Judge for Best of 16

Why and what was frozen. Simulated readers of an earlier draft (AI use statement) noted that the regime in which LLM judges rerank one model's samples had no LLM judge in the paper, and that the draft's practical sentence for coarse LLM judges, acc^*_{eq} picking a better judge than pooled correlation with eight labelled segments of WMT25, had been described after the fact. The fifth

protocol (`prereg/PROTOCOL_ppe_llm.md`) froze, before any judge rated a response of the PPE pool or measurement set, the judges, the prompt and the decoding, a completion rule, six rules (the paper's five and acc^*_{eq} read off the labelled prompts, its tie threshold chosen there), the draws (the 500 half-splits of the pool the reward-model panel was read on, so that every panel reads the same labelled prompts), the bootstrap (200 replicates resampling the pool's and the measurement set's prompts) and five hypotheses on the five sets' mean (Table 42). Its record and the scoring script were time-stamped at 15:11:06 UTC on 25 September 2026. The runner and the grader were written afterwards to its text, tested on the reward-model panel, on which the runner reproduces the paper's PPE picks exactly on four sets (on MATH the ten prompts Deviation 1 removes for panel B also leave that test's pool, and the eight-prompt accuracies differ by up to 0.7 points), and on synthetic judges, and frozen with the deviation below at 16:11:01 UTC, before either read a rating. The test outputs were deleted before that record was written, which is why it lists no result file; before the analysis the ratings were fixed by the line counts the record lists, not by hashes of their content, which the results manifest of 18:46:59 UTC gives.

The judges and the prompt. Twelve general-purpose LLMs served through one API, Alibaba Cloud's DashScope: Qwen3-8B, -14B, -32B and -235B-A22B with thinking off, Qwen3-30B-A3B- and Qwen3-235B-A22B-Instruct-2507, Qwen3-Next-80B-A3B-Instruct, DeepSeek-V3 and V3.2 and Kimi-K2-Instruct, all with open weights, and the closed Qwen-Flash and Qwen3.5-Flash. Each saw the released prompt and one response in a single user turn and was asked whether the response correctly and completely accomplishes the task, as one integer from 1 (certainly incorrect) to 10 (certainly correct); greedy decoding, at most eight output tokens, the first integer from 1 to 10 in the reply read as the rating, one reminder when there was none. This is one readout of an LLM judge among several: the mean of its rating distribution or a pairwise comparison breaks the ties a greedy integer leaves (Wang et al., 2025; Landesberg, 2026). Gemini and Gemma models, the smallest Qwen3 models and GLM-4.6 were tried in the pilot and left out before the freeze: the only Gemini key's credit was spent, the small models were not served to the account, and GLM-4.6 answered about one response a second, about six and a half hours for the 23,040 responses, which the completion deadline would have allowed; its pilot ratings, among the weakest, had been read.

What had been read. PPE's items and labels and the reward models' scores, read since the seventh draft (Section 6 and Appendix M); the fourth protocol's results, whose described acc^*_{eq} result is what L2 and L3 grade; and a pilot of 160 responses of ten prompts in neither the pool nor the measurement set, whose parse rates, rating distributions, pairwise accuracy and best-of-16 accuracy inside those ten prompts were read (the best-of-16 accuracy from 0.48 to 0.70 among the twelve judges). No judge had rated a response of the pool or of the measurement set. A PPE prompt is a question answered by one generator, and the same question can stand in two prompts with different generators: 100 of the 960 measurement prompts pose a question that also stands in the pool, and 4 of the ten pilot prompts a question that stands in the measurement set, answered by other generators. Without those 100, L2's and L5's point estimates are +0.3 and +0.6 points, against +0.3 and +0.5 with them.

Deviation 1. From about 15:50 UTC (the unfinished rating files' last writes) the API account answered every request with "Arrearage", its credit spent, in part by this scoring. Before the runner or the grader read a rating, a deviation, hashed and time-stamped with them, applied the completion rule then rather than at 03:00 UTC on 26 September and made the panel final: seven judges had rated at least 95% of both sets, the five with open weights Qwen3-32B, Qwen3-30B-A3B-2507 (by two responses), Qwen3-Next-80B, DeepSeek-V3 and DeepSeek-V3.2, and Qwen-Flash and Qwen3.5-Flash; Qwen3-8B, -14B, -235B, -235B-2507 and Kimi-K2 were dropped, their ratings of the measurement set 67% to 89% complete. The missing ratings of the kept judges are mostly the end of the item file, 765 MBPP-Plus responses of Qwen3-30B-A3B-2507 and 218 of Qwen3-Next-80B, besides one IFEval rating each of Qwen3-30B-A3B-2507 and Qwen-Flash, all imputed at 5.5 as frozen, which makes those judges pick at random on 47 and 12 MBPP-Plus measurement prompts and partly on 2 and 3 more; the same analysis on the 909 measurement prompts every kept judge rated in full is reported below and not graded. On panel B ten MATH prompts of the pool, holding a response a reward model has no score for, leave the pool, since acc^*_{eq} pools the pairs of prompts of equal size. The deviation's text gives "about 16:25 UTC" for its own time and "about 15:55" for the first refused request, estimates written before the freeze; its token's time is 16:11:01.

Results. Table 57 gives every rule with 8 and 24 labelled prompts and at half the pool on both panels, and Table 58 every judge. The graded tests, each with the protocol's interval and its percentile interval: L1 consistent (+0.4 points $[-6.4, +7.3]$, percentile $[-5.4, +8.4]$), L2 consistent (+0.3 $[-0.1, +0.7]$, $[-0.1, +0.7]$), L3 consistent (+0.5 $[-0.1, +1.0]$, $[-0.2, +1.0]$), L4 supported (-0.2 $[-1.2, +0.9]$, $[-1.2, +0.9]$), which could not come out not supported, pooling's regret at half being 2.0 points, and L5 not supported (+0.5 $[-0.1, +1.1]$, $[-0.0, +1.1]$); the percentile intervals change no grade, and by the protocol's rule the practical sentence stands as the grades stand. On panel B, L1 is supported (+4.9 $[+1.4, +8.3]$, $[+0.6, +6.9]$) and L2 consistent (+0.4 $[-0.4, +1.1]$, $[-0.4, +1.1]$).

Described, not graded. With eight labelled prompts on panel A every rule names the best judge of a set on 14% to 20% of draws, against 14% for a random pick, and the worse judge of a pair on 40.4% to 44.3% of the pairs whose accuracy on the measurement set differs; pooling's pick falls 2.8 points short of each set's best judge on average, 4.8 on MATH, where the judges span 10.9 points. L2's and L5's estimates come from MATH: without it they are +0.05 $[-0.34, +0.44]$ and +0.07 $[-0.47, +0.60]$, and without any other one set they stay between +0.30 and +0.42 and between +0.57 and +0.67. acc^*_{eq} pays an evaluator for tying responses whose labels tie, which best of 16 never rewards, and its statistic ties or is undefined on 4% to 28% of the decisive judge pairs, 15% on average, against at most 1.7% for the other rules, a tie resolved by the draw's priority as for every rule. In the protocol's label units, pooling's pick matches each rule's eight-prompt pick at about these budgets on panel A: r_W 8.9, the grouped mean ¡=8, the pairwise accuracy 22.2, best of 16 29.2 and acc^*_{eq} 20.4; on panel B: 27.0, 40.8, ¿48, 8.4 and 12.0. Pooling's curve rises little over the budgets (Figure 2a), so gains below a point map to many labels. At half the pool acc^*_{eq} orders the judge pairs worse than pooling on panel A by 7.8 points $[0.5, 15.0]$, whose percentile interval covers zero, and on panel B by 6.9 $[2.2, 11.5]$, whose percentile interval does not. Of panel B's advantage at half, +4.9 points over all pairs, 4.1 come from pairs of an LLM judge and a reward model or the length baseline, 0.6 from pairs of two reward models or a reward model and the length baseline, and 0.2 from pairs of two LLM judges, among which the grouped mean's advantage is +0.7 points (`scripts/same_model/ppe_llm_described_r18.py`). The label-free mean of the seven judges' ratings, a candidate no rule could name, reaches 65.3% best-of-16 accuracy, 1.0 to 5.5 points above the best judge of each set and 3.4 on average, percentile $[1.2, 4.6]$ over 2,000 resamples of the measurement prompts with the best judge re-selected in each (`scripts/same_model/ppe_llm_ensemble_r18.py`); their median reaches 61.1%, and the best reward model of each set beats the mean on 2 of the five (Table 59). On the 909 measurement prompts every kept judge rated in full, acc^*_{eq}'s gain over pooling with eight labelled prompts is +0.3 $[-0.1, +0.7]$ points and the grouped mean's ordering advantage at half -2.3 $[-9.2, +4.6]$, which the grader would call not supported. Every judge rated at least 98.8% of each of MMLU-Pro, MATH and GPQA before the account ran out, the rest imputed; on those three sets, read with the frozen runner's functions and draws (`scripts/same_model/ppe_llm_all12.py`, point estimates only), the twelve judges' best-of-16 accuracy runs from 50.3 (Qwen3-8B) to 60.9 (Qwen3-235B-2507), both dropped by the deviation, pooling's pick falls 3.3 points short of each set's best judge, and against pooling's pick acc^*_{eq}'s gains +0.8 points, the grouped mean's -1.5, the pairwise accuracy's -0.8 and best of 16's on the labelled prompts -0.4, where the seven kept judges give +0.5, -0.4, +0.3 and +0.7: the last two change sign with the panel, L5's quantity among them.

Addendum: the judges' mean on WMT25. Before the text offered the mean of the judges as advice, an addendum to the protocol (`prereg/PROTOCOL_ensemble_wmt25.md`), written after the grades above with the mean's PPE result known, froze a test of it in a second regime: WMT25's thirteen LLM judges on the fourth protocol's two LLM-judge sets, translations of eight systems against MQM and ESA labels. The mean is taken over the judges' scores standardised within a configuration and reads no label; each rule is paid best-of-n recovery on the evaluation half of the fourth protocol's 300 draws, which the script replays and checks against that protocol's frozen picks. E1: the mean recovers more than pooling's pick with eight labelled segments; E2: more than the in-sample best judge, an oracle that reads every label. The intervals resample documents, 200 replicates of 40 half-splits. Its rule: only if E1 is supported on both sets may the text say that averaging beat choosing on WMT25 as on PPE. Its record, script and imported modules were time-stamped at 20:23:21 UTC on 25 September 2026, when no combination of WMT25's judges had been computed; the replay had been checked with one judge in the mean's place, whose results the fourth protocol holds.

On MQM, E1 is supported (+15.7 $[+6.3, +25.1]$, percentile $[+6.7, +25.5]$) and E2 consistent (+1.4 $[-9.1, +11.9]$, $[-11.7, +9.1]$); on ESA both are contradicted, E1 at -4.5 $[-7.8, -1.2]$, whose percentile interval $[-6.7, +0.02]$ just reaches zero and would call it not supported, and E2 at -13.0 $[-16.2, -9.7]$, $[-16.3, -10.1]$ (Table 60). By its rule the sentence about combining judges is stated for PPE alone. Described after the grading: on ESA the mean beats pooling's pick on 8 of the 13 language pairs and the in-sample best judge on 3; on en-ar and en-bho, where 8 or more of the thirteen judges pick worse than chance over all segments, its recovery is -18.7 to -11.9 points, worse than chance too, while pooling's pick with eight labelled segments recovers at least 31.5. Without those two pairs, a split chosen after the grades were read, E1 on ESA is +4.4 points and E2 -4.1. A label-free mean follows the majority of its panel, and where most judges are wrong a few labels are what tell the others apart (`scripts/decision/ensemble_wmt25_described_r18.py`).

Second addendum: a jury of three. A second addendum (`prereg/PROTOCOL_jury.md`), written after the first was graded and time-stamped at 20:59:12 UTC on 25 September 2026, when no combination of fewer than all of a panel's judges had been computed on any set, froze a use of the eight labels that keeps what each regime rewarded: the jury of the three judges that pooled correlation names in turn on the eight labelled units, the first being pooling's own pick, their scores averaged as each regime averages all its judges (raw ratings on PPE, standardised scores on WMT25). It is paid on the protocols' own draws against pooling's single pick (J1) and against the mean of all the judges (J2), on PPE's five sets with the seven judges and on WMT25's two LLM-judge sets with the thirteen, with the fifth protocol's bootstrap on PPE and the first addendum's on WMT25. Three, the size of the jury of Verga et al. (2024), was fixed and not tuned, and no weighting was used. The replay reproduces pooling's picks in the frozen runs, and on PPE the fifth protocol's bootstrap replicates, exactly. Its rule: if J1 is supported on all three sets, the text may advise averaging the three judges pooled correlation ranks highest rather than using the one it ranks highest, and if J2 is too, that this also beats averaging every judge.

J1 is supported on all three sets: +3.8 $[+3.0, +4.6]$ points of best-of-16 accuracy on PPE (percentile $[+3.0, +4.7]$), +14.0 $[+8.1, +20.0]$ points of recovery on MQM ($[+8.3, +20.1]$) and +7.7 $[+6.3, +9.2]$ on ESA ($[+6.6, +9.5]$), the jury ahead on all 5 PPE sets, both MQM pairs and 12 of the 13 ESA pairs. J2 is contradicted on PPE (-2.4 $[-3.6, -1.3]$), not supported on MQM (-1.7 $[-11.0, +7.7]$) and supported on ESA (+12.2 $[+9.2, +15.2]$): where most judges are informative the mean of all of them is as good or better, and where most are not the jury keeps what the labels found (Table 61). By its rule the text advises the jury over the single pick and not over the mean of all. Described after the grading: juries of two, four and five reach 61.5, 63.7 and 64.3% on PPE, 41.6, 46.9 and 47.4 on MQM and 39.1, 40.9 and 40.8 on ESA, against 62.8, 45.5 and 40.6 for three, so that on ESA every jury from three to five beats the mean of all and on PPE the larger ones approach it; ranked by the grouped mean the jury of three reaches 62.8, 50.1 and 40.8; and at half the pool or half the segments it beats pooling's pick by +3.4, +9.3 and +5.6 points. It costs three judge calls a response, against one for a single judge and seven or thirteen for the mean of all.

Three judges at random. Read after both addenda and graded by none: `scripts/decision/jury_baselines_r22.py` replays the second addendum's draws and replicates exactly, its pooling pick, jury and J1 reproducing `results/jury_r18.json`, and pays beside them a jury drawn at random, the mean over every set of three judges of the panel of that three's payoff, which costs the jury's three calls a response and reads no label (Table 62). On PPE the jury leads it by +0.2 $[-0.3, +0.7]$ points; on WMT25 by +8.8 $[+2.8, +14.9]$ on MQM and +19.8 $[+17.9, +21.8]$ on ESA, where a random three falls below pooling's single pick on 10 of 13 language pairs. Without en-ar and en-bho, where most judges pick worse than chance, the ESA lead is +10.9 points, the mean of the other 11 language pairs, positive on 11 of them (point estimates, read after every protocol), so the jury does more there than drop the judges below chance. On the 860 of PPE's 960 measurement prompts whose question does not also stand in the pool, the jury against pooling's pick, the jury against a random three and a random three against pooling's pick are +3.7, +0.2 and +3.6 points, against +3.8, +0.2 and +3.6 on every prompt.

W LIMITATIONS IN FULL

The exact decomposition and Corollary 1 are statements about Pearson correlation; under the rank and per-row protocols that several benchmarks use, the bound on cluster-constant predictors is the different quantity of Propositions 4 and 5, and the corollary's band has no counterpart there. Under a pooled rank coefficient that bound is an upper bound we certify rather than a maximum we compute, so the Bound column of Table 72 overstates by an unknown amount on those 9 rows, by at most 0.0623 above the best predictor our search finds. A second gap is between the within-cluster coefficient and the decisions it stands for, which Section 3 states and this quantifies. r_W weights a cluster by how much the predictions vary in it, so an evaluator whose output scale differs by cluster moves it without changing any decision: `scripts/decision/rw_invariance.py` multiplies each cluster's predictions by a random positive factor with a unit log standard deviation, which leaves every within-cluster order and so every decision unchanged, and the mean of the per-cluster correlations, which a per-row protocol computes and which is invariant in exact arithmetic, moves by at most 4×10^{-14} while r_W moves freely. No reading of Section 5 takes any of these coefficients for a sufficient statistic of a decision; where the scale of an evaluator's predictions varies by cluster for reasons of its own the grouped mean is unmoved and r_W is not, which is one reason the reporting protocol asks for both, and Section 6 measures what the grouped mean gives up when the labelled clusters or their candidates are few. The exploratory evidence is six benchmarks, and 19 configurations of them whose clusters are sets somebody chooses from, to which the frozen validation adds two and the same-model checks two more, which is a small number of independent units for a claim about evaluator benchmarks in general. Six units also decide what the inference can be: a benchmark-resampled interval behaves as an 85% interval and a test at about 15% rather than 5% (Appendix Z), so that resample is kept in the release and read as spread, and no claim across benchmarks is made as a test: what is reported is the direction on every benchmark and every task family, the effect sizes, and the leave-one-out and family-level readings, with the sign test's floor of $1/64$ stated for what it is, since its units are dependent and the comparison it would test was named after a subset of the data had been seen (Appendix I). The label-only screen is not a threshold at a share of one half: on the primary cells every cell above the line belongs to one of 3 benchmarks and every cell below to one of 3, with 0 on both sides, so that contrast compares benchmarks wearing the clothes of a comparison of cells, and the ordering over the labels of one benchmark, which is free of that confound, is absent, running from -0.60 to $+0.23$ on the 3 benchmarks whose labels give it spread. What replaces it is in Appendix I: the share manipulated with everything else fixed, in simulation and inside each benchmark, moves pooling's excess selection error over evaluator pairs with it, and that excess is positive at every share either manipulation produces, which is why the share is reported as the stake; it is a statement about the ordering of evaluator pairs, which Section 5 separates from the utility of the evaluator a rule selects.

The released panels do not cover the decision that motivates the within-cluster component most directly, choosing among several samples of one model for one prompt. Appendix I supplies the nearest condition the released panels allow, a candidate set whose systems are matched on quality, and this states what the corpora hold. No benchmark in our census releases evaluator scores on a same-model set. The closest is the axis-rating release of Stiennon et al. (2020), audited with its CNN/DailyMail part added to the Reddit TL;DR part the census uses, by `scripts/decision/same_model_audit.py`: 1,677 prompts, 24 policies, a median of 8 candidates a prompt and up to 21, but only 22 of its 13,288 prompt-by-policy cells hold more than one sample of the same policy and 0 hold three. What it does hold is 9 policies that are one model at two decoding temperatures, and 270 prompts carrying both variants of one of them, which is a same-model choice between two candidates with human ratings on four axes. It carries no evaluator score for any of it, which is why the four labels of its TL;DR part are the whole of the exclusion in Appendix H. The same-model checks the paper does make are Appendix L, on PRM800K with a panel of reward models we chose, and Appendix M, on the PPE sets with the panel their authors chose, both under designs fixed before scoring; each is a different object from the released panels used everywhere else here, one generator per prompt, verifiable tasks, and five or six evaluators. The nearest published evidence, a 24-prompt same-model pilot scored against a stronger model rather than against human labels (Landesberg, 2026), points the same way and is smaller. Dropping a whole field rather than one benchmark, which is the coarser check because benchmarks of one field share source corpora and evaluator families, leaves the gap in selection error between the pooled and the within-cluster coefficient at 3.4 to 5.8 points. The resampling unit differs by claim and so

does the population each claim speaks about, which is worth stating plainly. A cluster bootstrap on one benchmark speaks about fresh clusters of that benchmark's kind, new speakers or new source segments, with its systems held fixed; the two-way resample of Appendix AB adds fresh systems drawn from the same pool; and the benchmark resample kept in the release (Appendix Z) speaks about a fresh draw from the six benchmarks we hold and not from evaluator benchmarks at large, which is why the field-level leave-one-out is reported beside it. None of the three speaks about fresh evaluators: the evaluators are the released ones throughout. Four things are stated here rather than in the main text, for room. Running the decision score itself on the selection half is the reference rule of Table 65; whether it can stand in for a coefficient that several decisions read is measured in Appendix I and depends on which decision statistic it is, and it is not uniformly the better rule even under its own utility, since a decision score computed on half the clusters is noisier than a correlation over all their items, and on the primary best-of-n cells it errs on 25.0% against 22.2% for the grouped mean. The oracle of Proposition 3 is the limiting case of the two components pulling apart: it reaches 0.816 pooled on speechocean762, above every released evaluator but the strongest reproduced system, while ordering a learner's utterances backwards at a pairwise accuracy of 0.000. Splitting the held-out pairs by what the choice was worth answers the worry that a rule wins only where the evaluators are near-equivalent and a flip costs nothing; it is the other way round. On the 8% of primary pairs whose held-out decision scores differ by less than 0.01 every rule is near chance, 46.5% for the pooled coefficient against 42.6% for the grouped mean; on the 26% between 0.01 and 0.05, 33.8% against 24.5%; and on the 67% at or above 0.05, 14.8% against 9.3%. In the pair-level experiment of Appendix H, which scores the choice between two evaluators and not the evaluator selected from a panel, the payoff of grouping in decision units is small, +0.0030 $[+0.0008, +0.0082]$, and is better read against what any rule could buy: on a scale where choosing at random realises no gain and choosing with the answer realises all of it, the pooled coefficient recovers 72% of the headroom, the within-cluster one 80% and the grouped mean 85%. The system-clustered cells leave more room between the rules than the primary ones, 31.5% against 28.9%. Measured on the clusters the coefficients were read from, the within-cluster coefficient disagrees with the decision on 17.2% of the 8682 evaluator-pair comparisons against 22.0% for the pooled one, a larger gap than the held-out one, and resampling whole benchmarks puts it at 2.7 to 8.9 percentage points; on RealSumm, choosing inside one system, the benchmark's own per-document rule errs on 51.6% of pairs where the pooled coefficient errs on 27.6%. The equivalence margin of the judge swap (Appendix E) is defined from a quantity the contrast it bounds does not involve, the spread of the five experts around their own panel, but it was fixed after that contrast had been computed and not pre-registered, so the smallest margin the data support is reported beside it and a reader who prefers another margin can apply it. The population ceiling assumes items exchangeable within a cluster, which a corpus whose speakers read different sentences satisfies only approximately: on speechocean762 almost every sentence is read by one speaker, so speaker identity and the twenty sentences that learner read cannot be separated. Within-cluster reliability needs rater-level labels, and where rater identities are not released the one-way estimate charges rater leniency to error, which can drive it below zero. Per-item predictions of published systems are rarely released, so the reproduced systems are those with public checkpoints or training code, and their checkpoint-selection regimes differ, which Table 70 records. History pooling and cluster-mean conditioning need several items per cluster at evaluation time, which isolated items do not provide. The commercial services are scored through interfaces whose versions a client cannot pin (Appendix X), and the analysis is of utterance-level scores only: the word and phone levels that pronunciation feedback uses are a three-level design that the two-level framework here does not cover.

The census. Every numeric human label of every normalised benchmark is crossed with every column of its item table that names a unit the benchmark samples, which gives 201 configurations over the 16 benchmarks and parts that Table 63 lists with their fields, sources and counts; 138 have at least five clusters and a typical cluster of more than one item, and enter the summaries. 21 have a between-cluster label share above one half, 7 of them with two items per cluster, where the in-sample share tends to $(1 + \rho)/2$ as clusters multiply and is therefore expected above a half whenever $\rho > 0$ and the clusters are many, though not in every sample and not with few clusters (Appendix B). Taking the intraclass correlation instead of the in-sample share puts 13 above the line. Both readings are right about different things: the in-sample share is what a pooled coefficient computed on the released labels actually mixes, and the intraclass correlation is what a larger sample from the same population would show. One card per configuration ships with the release.

X How the Commercial Services Were Called

All three were called on 13 September 2026 through their public interfaces, from one machine whose clock is UTC, on the audio and reference text the benchmarks release and nothing else: Azure between 18:10 and 18:18, iFlytek until 18:39 and the judge between 19:48 and 19:53 UTC, which are the write times of the request logs. No speaker identifier, no label and no history was sent. Correlations are scale free, so no rescaling was applied to any vendor's output. Each arm's log, one line per request, holds the request parameters below, the fields taken from the response and the raw values; the judge's logs ship under `results/api/` and the two speech services' are withheld, for the reasons the ethics statement gives, with the aggregates computed from them shipping instead; the interface version behind a product name cannot be pinned by a client, so the logs and the dates are the record of what was measured.

Microsoft Azure AI Speech, pronunciation assessment. The REST recognition endpoint `/speech/recognition/conversation/cognitiveservices/v1` with `language=en-US` and `format=detailed`, and the assessment parameters `GradingSystem=HundredMark`, `Granularity=Phoneme`, `Dimension=Comprehensive`, `EnableMiscue=true` and `EnableProsodyAssessment=true`. The audio is the released 16 kHz mono file, unmodified. The response is JSON; we take `NBest[0].PronunciationAssessment.{AccuracyScore, FluencyScore, ProsodyScore, CompletenessScore, PronScore}` and the recognised text `NBest[0].Display`, and record them as the accuracy, fluency, prosody, completeness and total fields of the log. All 2500 utterances returned a result; five carried no prosody score, so the prosody aspect is reported only where the service supplies one. The service is deterministic: 500 repeated requests returned byte-identical scores.

iFlytek, spoken language evaluation. The streaming endpoint `wss://ise-api.xfyun.cn/v2/open-ise` with `ent=en_vip`, `category=read_sentence`, `rst=entirety` and `check_type=common`, audio sent as raw 16 kHz frames. Its request carries a declared speaker group, which we set to `pupil` for speakers aged twelve or under and `adult` otherwise; the English engine rejects the third value the interface documents. Scores are the accuracy, fluency, integrity and total fields of the returned XML, on a nought to five scale. The service is not deterministic: repeated identical requests move the total score by a quantised ± 0.25. Its trial allowance ran out after 519 utterances over 29 speakers, which is why the label-free arms were run on Azure and the identity arm on a judge instead. Which utterances those are is a fact about the order of the requests and not a random draw: the requests walked the released test list, which is stored in contiguous speaker blocks, with several workers in parallel, so the subset holds utterances 1 to 878 of the list, 25 speakers complete and 4 in part. `scripts/api/subset_control.py` scores every other evaluator on exactly those 519 utterances so that iFlytek can be read on one item set: on the total, within speakers, iFlytek reaches 0.410 there against 0.439 for Azure, 0.440 for HMamba and 0.357 for the WavLM regression, pooled 0.658, 0.612, 0.707 and 0.667, which places it 3 of 7 within speakers on its own subset.

Gemini 3.6 Flash. The chat completions interface of the Gemini API at temperature zero, requesting the model string `gemini-3.6-flash` through an OpenAI-compatible endpoint, one summary per request as a single user message, the answer parsed as the first digit between one and five; a request that returns no parsable digit after four attempts is recorded with its error and excluded from that arm, and the logs record 40 such exclusions over the 2190 requests of the four arms. The log records the document, the system, the condition and the parsed score. The client libraries were `requests` 2.34.2 for Azure and the judge, `websocket-client` 1.9.2 for iFlytek and `defusedxml` 0.7.1 for its XML, from one Python 3 environment whose full lock file ships with the release. The prompt paraphrases the coherence definition of Fabbri et al. (2021), and one sentence of provenance optionally precedes it. The prompt in full:

```
[provenance sentence, in the two swap arms only]
Rate the coherence of the summary on a scale of 1 to 5, where
coherence is the collective quality of all sentences: whether the
summary is well structured and well organized rather than a heap
```

```
of related information. Answer with a single digit and nothing
else.

Summary:
<summary text>
```

with the provenance sentence “The summary below was produced by a state-of-the-art summarization system.” or “The summary below was produced by an early, low-quality summarization system.”, attached to the strong and the weak systems respectively in the aligned arm and the other way round in the reversed arm.

The comparison against one human expert, in full. The released label is a floored median of five raters, so a service scored against it is not being compared with one expert on equal terms. The human row instead scores one rater against the unrounded median of the other four, and the service and the reproduced system are scored against the same target. Within speakers on the total the service reaches 0.530, a single expert 0.507 and the reproduced system 0.511. The equivalence margin is the range of the five experts around their own panel, a quantity the contrast it bounds does not involve. Equivalence holds on fluency, where the 90% interval $[-0.042, +0.034]$ lies inside that label’s margin of 0.123 with both one-sided tests at $p = 0.0005$, the floor 2000 resamples can resolve, and on no other label; on the total the data are consistent with any difference below 0.056 and establish nothing smaller, against a smallest detectable difference of 0.063. The one difference that resolves runs against the service, -0.061 $[-0.116, -0.008]$ on accuracy, and does not survive a correction over four labels at 0.088. The reproduced system is above the expert on all four labels, so agreement at that level is not particular to the service.

The four arms of the provenance swap. The between-system correlation of the judge is 0.944 when the declared provenance is aligned with the systems’ true standing and 0.817 when it is reversed, against 0.934 with no provenance sentence and 0.936 on a repeat of that same request, which moves 35% of the item scores and is the judge’s own jitter. Aligned minus reversed is the swap contrast and repeat minus plain is the noise contrast; the judge swap (Appendix E) differences the two on the same resampled systems rather than testing each against zero.

The interface version behind each name is not something a client can pin, so these numbers are a measurement of what those services returned on those dates and not a property of a fixed artifact. That is a limitation of measuring deployed systems at all, and it is the reason the per-item outputs are kept: the judge’s ship with the supplement, and the two speech services’ are withheld from it, for the reasons the ethics statement gives.

Y Benchmarks and Data Processing

speechocean762. The corpus is released under CC BY 4.0 through OpenSLR resource 101 (Zhang et al., 2021). The official split has 125 training and 125 test speakers with no overlap, and every speaker reads 20 sentences. Of 4947 distinct sentence texts, 4894 are read by a single speaker, so sentence identity carries almost no information across speakers and the text-only control uses surface features instead. Five experts scored every utterance, but the release does not identify experts across utterances, so the rater position is used as a rater index. That is an assumption, and the fixed-panel estimator of Appendix D depends on it, so we tested it rather than asserting it. After the utterance main effect is removed, position carries a main effect of $F = 601$ to 2750 on $(4, 9996)$ degrees of freedom across the four labels, with partial η^2 from 0.19 to 0.52; the profile of five position offsets computed on one half of the speakers reproduces the profile computed on the disjoint other half with a correlation of at least 0.94 in the worst of 200 splits. An offset pattern stable across disjoint speaker sets is what a fixed panel means operationally. The alternative explanation, that the five scores are simply stored in sorted order, is refuted three ways: the position means are not monotone in any label, at most 16% of rows are monotone where sorted storage would force all of them, and once the position offsets are subtracted the monotone count falls to chance. The assumption is not cosmetic and it moves the estimate in the direction that favours us: permuting the five scores within each utterance, over 1000 permutations, takes the within-speaker reliability of fluency from 0.744 to -0.033 and of prosody from 0.704 to -0.113, so without a fixed panel the within-speaker human ceiling on those two labels would be zero. What the check establishes is that the column factor is

fixed and consistent, which is what the crossed model requires; it cannot establish that a position maps to a named person rather than to a fixed scoring stage. It is a check of a modelling assumption and not a substitute for rater metadata, which the release does not carry and which we have not obtained from the corpus authors, so both reliability estimators are reported wherever the choice matters. The release describes the aggregation only as the average or median of the expert scores. We recovered the exact rule and verified it on all 5000 utterances: accuracy, fluency and prosody are the median of the five expert scores, completeness is ten times the median of the five completeness fractions, and the total is the median of the five expert totals rounded down. Completeness equals its maximum for 99.5% of utterances, and the non-maximal test utterances belong to three speakers, so completeness is excluded from the main analyses. The test utterances are stored in contiguous speaker blocks, and any contiguous slice of the test list changes the set of speakers. WavLM-large features are hidden states of layers 6, 12, 18 and 24 averaged over time and concatenated, computed in half precision from 16 kHz audio. Speaker-verification embeddings come from the CAM++ model (Wang et al., 2023a).

MOS benchmarks. How far the shortcut moves inside one task is visible across the three. On BVCC the between-system share is 77% and the oracle reaches 0.782 in utterance-level Spearman correlation against 0.897 for UTMOS22 (Saeki et al., 2022), whose public strong learner has a within-system correlation of 0.541 where listener agreement allows 0.799; on SOMOS (Maniati et al., 2022) the oracle reaches 0.479 against 0.681 (Vioni et al., 2023), and on VCC2018 (Lorenzo-Trueba et al., 2018) 0.476 against 0.718 (Tseng et al., 2021).

BVCC (Cooper & Yamagishi, 2021) is used through the VoiceMOS 2022 main-track release (Huang et al., 2022a), whose archive ships 3382 of the 7106 audio files. The remaining 3724 files come from Blizzard Challenge releases that the organisers may not redistribute, and we reconstructed them with the release's own download, gather and preprocess scripts, verifying every source archive against the checksums in those scripts. Mean listener ratings equal the official mean opinion scores for all 7106 utterances. Our port of the SSL-MOS baseline (Cooper et al., 2022) agrees with the organisers' released predictions to a mean absolute difference of 0.001 on the development split, and 0.002 on the utterances we reconstructed. SOMOS (Maniati et al., 2022) is used through the official split-1 clean subset, whose labels and rating filter we reproduced exactly from the released raw scores. VCC2018 (Lorenzo-Trueba et al., 2018) has no official per-utterance label, so we use the listener mean, which is the convention of the MOS prediction literature (Lo et al., 2019), and the 4000-utterance test split of the open-source MBNet implementation, as redistributed with LDNet (Huang et al., 2022b) and used by Tseng et al. (2021). TTSDS2 (Minixhofer et al., 2026) is used with its released listening-test ratings, and its per-domain system means reproduce the published table for all 20 systems.

Licences. As their releases state them: speechocean762, VCC2018, RealSumm and HelpSteer2 CC BY 4.0; BVCC BSD 3-Clause; SOMOS CC BY-NC-SA 4.0, whose labels are therefore withheld; SummEval, FRANK, HANNA, TTSDS2 and PRM800K MIT; TL;DR axis ratings a modified MIT licence; mt-metrics-eval Apache 2.0, with no separate licence for the scores; RewardBench 2 ODC-BY. PPE's best-of-K sets license their prompts under MIT (MATH, MMLU-Pro), CC BY 4.0 (GPQA) or Apache-2.0 (IFEval, MBPP-Plus) and leave the responses to the terms of use of the models' providers, and GPQA's authors ask that its questions not be shown in plain text or images online; the supplement therefore ships PPE's items without their text, each prompt and response replaced by its SHA-256, and the scripts that rebuild the files byte for byte from the public release. The release's cards (`scripts/release/licences.py`) record the evidence for each census benchmark.

Text and translation benchmarks. WMT22 MQM (Freitag et al., 2022) is used for en-de, zh-en and en-ru with the official segment scores and the metric submissions distributed by the shared task. Our reconstruction from the released error spans reproduces the official scores exactly for en-de and zh-en; for en-ru 18,897 of 21,040 segments match, 1853 carry no span rating and 290 differ, and the official scores are used throughout. Human reference translations and the systems the shared task marks as outliers are excluded, which is the configuration under which the published segment-level Kendall correlations of its Table 11 reproduce to three decimals. WMT23, WMT24 and WMT25 are read from the mt-metrics-eval v2 archive of the WMT metrics tasks (Freitag et al., 2023; 2024; Lavie et al., 2025), whose package is released under the Apache License 2.0 and whose archive states no separate licence for the scores; the supplement carries only statistics derived from them

and the adapters' cards, not the scores or the translations. HANNA (Chhun et al., 2022) is analysed without the human-written stories, as in its own story-level protocol, and its evaluators include the LLM ratings released with Chhun et al. (2024). RealSumm (Bhandari et al., 2020) uses the LitePyramid recall labels, which we reproduced exactly for the 2400 summaries that carry an official label. FRANK (Pagnoni et al., 2021) uses the released factuality scores; its official aggregate cannot be rebuilt exactly from the released sentence annotations, and the paper's majority rule matches 2141 of 2246 summaries, which we record rather than repair. HelpSteer2 (Wang et al., 2024c) and the TL;DR axis ratings (Stiennon et al., 2020) have no official evaluator protocol and enter only the label-only analysis; of the latter we use the Reddit TL;DR subset, which is the eponymous one, rather than the CNN/DailyMail axis ratings released alongside it. PRM800K (Lightman et al., 2024) is released under the MIT licence, with the MATH problems it annotates (Hendrycks et al., 2021); its phase-2 test split is used whole, the step ratings read from the release's label records and the final answers graded with the release's own grader, as Appendix L states, and it is not part of the census, since the only evaluator score it carries is that of the reward model that chose which solutions to label, one evaluator where the decision experiment asks for at least four, and its use here is the panel we scored.

SummEval. The annotations are released under the MIT license (Fabbri et al., 2021). One hundred CNN/DailyMail test documents are each summarized by 16 systems, and three experts rated every summary for coherence, consistency, fluency and relevance on a five-point scale. The label is the mean of the three expert ratings, which matches the labels used by G-Eval, UniEval and BARTScore on all 1600 summaries. Expert identities are not documented across summaries, but the rating table carries a fixed three-column structure and the position test passes on it, so the rule of Appendix D licenses the fixed-panel estimator here. We print the one-way estimate in the main text as the lower of the two and give both in Appendix D. The crowd ratings in the same release are the negative control that calibrates the test, and they fail it. Evaluator scores are taken from the released outputs of G-Eval with GPT-4 (Liu et al., 2023b), UniEval (Zhong et al., 2022), the ChatGPT evaluator of Wang et al. (2023b), and the metric scores distributed with BARTScore (Yuan et al., 2021). G-Eval scores are the mean of the parsed scores of the 20 sampled responses, which reproduces the published per-document correlations within 0.02 for coherence, consistency and relevance and within 0.06 for fluency.

Z Power

When items of a cluster are correlated, the standard error of a correlation difference shrinks with the number of clusters rather than the number of items (Table 71). A paired bootstrap resamples both evaluators together, so the minimum detectable difference belongs to the pair and not to the benchmark: on the speechocean762 total score it is 0.026 within speakers for one model under two training seeds and 0.201 for two unrelated families, and over the 6 pairs of reproduced published systems it runs from 0.061 to 0.201, median 0.094.

What a resample of six benchmarks is worth. The intervals of the main text resample clusters, documents on WMT, within each corpus and jointly across a benchmark's configurations (Section 4), and their coverage for the paired utility difference is measured in Appendix AA; a bootstrap over whole benchmarks, the coarser unit, since cells of one benchmark share labels, items and evaluator families, is kept in the release as a second reading, and its level is checked here because six is a small number of units. Simulating that procedure: benchmark effects drawn around a true mean, cells inside each benchmark carrying their own noise, and the same nominal 95% percentile interval over resampled benchmarks. At six benchmarks it covers 85% of the time and excludes zero on 15% of draws where the true mean is zero, so it behaves as an 85% interval and a test at roughly 15% rather than 5%; at twelve benchmarks the figures are 90% and 10%, and at twenty-four 93% and 7%. Drawing the benchmark effects from a skewed distribution instead leaves them at 84% and 16%. The one-sided sign test over benchmarks rejects on 2% of null draws at six benchmarks, because at six units it can only reject when all six agree, which costs it power: on draws with a true mean of three points it rejects 28% of the time against 66% for the interval. That is the trade this paper makes. Intervals over benchmarks are reported as spread, and the statements that have to survive are the ones the sign test carries. The sign test's own null is that each benchmark's sign is a fair coin, which needs the units to be independent, and the three summarization benchmarks share a source corpus and evaluator families; the family-level reading of Appendix I is the coarser unit, and neither reading

quantifies the mean cost over evaluator benchmarks at large, which six units cannot do. The paired contrasts between two evaluators on one benchmark, which Appendices E and AB read, have their level checked in the same way: two evaluators drawn with identical population parameters, the paired cluster-bootstrap interval of their difference taken as `clustereval.paired_delta` takes it, and the share of nominal 95% intervals excluding zero counted over 300 draws. That share is 11.3% for r_T and 11.3% for r_W at nine clusters, 8.3% and 9.0% at sixteen, 5.7% and 7.7% at twenty, 9.0% and 6.0% at one hundred clusters of sixteen, which is the SummEval design, and 5.3% and 7.7% at 125 clusters of twenty, the speechocean762 design. Those are the levels at which the paired contrasts of Appendices E and AB are to be read, and they are read as description: a Bonferroni correction over a table needs the base test valid at the corrected threshold, which this simulation checks only at the nominal one, so no contrast in the paper is called significant after correction, and the release's multiplicity ledger records how many paired intervals exclude zero at the nominal and at the corrected thresholds without a claim about the level of the latter.

AA Coverage of the Utility Interval

The interval of Table 5 is the estimate ± 1.96 standard deviations over 200 corpus bootstraps of 300 splits each, a replicate reading b distinct clusters of the resampled corpus weighted by their multiplicities and, at half, a random half of those present; and the coverage read off simulations of correlation differences (Appendix Z) does not go through what the estimate goes through: a finite draw of labelled clusters, the argmax that names an evaluator with a tie-break, a ratio scored on the other half, and a mean over configurations. `scripts/decision/utility_calibration.py` runs the pipeline itself, with the formulas of the main run checked against it on random weights, inside worlds whose truth is known or computable, 200 replications of the whole pipeline per world and configuration (Table 73). In the *real* world the configuration's units, clusters or documents, are the population and a corpus is as many units drawn with replacement, copies being distinct clusters of the new corpus; its truth is the mean over such corpora of each corpus's own estimand, 300 corpora deep; the engine's own draw of 300 splits on the released corpus, with its own streams, puts the estimand at $+1.4$ points at half and $+0.9$ at eight over the primary configurations against the main run's $+1.6$ and $+1.2$, two independent draws whose difference at half is within two Monte Carlo standard errors of a split mean (Appendix J). That world has atoms, so a cluster can sit on both halves of a split, which a fresh benchmark would not arrange; it is the sampling model the bootstrap itself assumes, and the check is of the procedure under that model. A copy is treated two ways at two levels: the procedure treats the world's corpus as it would any benchmark, each unit a cluster of its own, while the interval's bootstrap, inside each replication as in the main run, keeps every copy of a cluster it resamples on one side of a split, weighted by its multiplicity. In the *null* world the labels and clusters are the configuration's, the panel is replaced by evaluators that are the label plus noise of one variance, matched to the panel's median pooled correlation, and the noise is fresh in every corpus, so the evaluators are exchangeable and the true paired difference is zero exactly, whatever the finite draws do; the *small effect* world scales one evaluator's noise by 0.9, and the *uneven sizes* world cuts a random half of the clusters to two items, so that few clusters carry the payoff's denominators. None of those panels holds what pooling and grouping disagree about, an evaluator that ranks the clusters well and the candidates poorly beside one that does the reverse, so the *cross* world builds the panel from the label's between- and within-cluster parts, y^B and y^W: one evaluator is $y^B + 0.3\,y^W$ plus noise, another $0.3\,y^B + y^W$ plus noise, the rest the label plus noise of 2 times the standard deviation, so that the two are the panel's best; its truth, by Monte Carlo, runs from -4.0 to $+38.9$ points at half the clusters: $+32.8$ to $+38.9$ on the four speechocean762 configurations and $+15.7$ on FRANK, where the labels' between-cluster share is 0.66 to 0.75 and pooling selects the evaluator that ranks the clusters, within a point of zero on HANNA and RealSumm, and -2.7 and -4.0 on SummEval fluency and consistency, whose shares, 0.09 and 0.05, are among the smallest; the interval covers it on 98% of replications, 94% to 96% on speechocean762 and 88% on FRANK, the lowest, and 98% at eight. Over the primary configurations the interval covers the truth on 94% of replications at half the clusters (80 to 100% by configuration over every set) and 96% at eight; in the null world on 99% and 100%, excluding zero on 1.1% and 0.4% of replications where the truth is zero; with a small effect on 99% and 100%, and with uneven sizes on 99% and 100%; the interval of the mean over the primary configurations, its replicates resampled jointly within a benchmark, covers its truth on 93% at half and 98% at eight, the mean over the five RewardBench 2 configurations on 95% and over the three WMT23 configurations on 95%. The graded sets hold five and three configurations, so the null

world's mean over random subsets of five configurations excludes zero on 1% of replications at half and 0% at eight, and over subsets of three on 1% and 0%. What the intervals are, by world and budget. At half the clusters in the real world they are nominal to short: the interval of the primary mean has a bootstrap standard deviation 0.95 times the true one and covers on 93%; per configuration the ratio runs from 0.73 on speechocean762 fluency to 1.53 on RewardBench 2 focus over the configurations on which the two rules disagree on at least a twentieth of the draws, below one on 9 of the 16 such primary configurations, and the three configurations whose true gains are largest in size, of either sign, are covered on 81% (HANNA relevance, ratio 0.84), 80% (SummEval coherence) and 87% (FRANK), the paired difference being skewed where it is large; the configurations left out of the range, with ratios of 2.1 to 19, are those on which the two rules almost never disagree, where the estimate barely moves. With the frozen bootstrap of 40 splits per resampled corpus the intervals at eight clusters were conservative, the primary mean's bootstrap standard deviation 1.47 times the true one and the per-configuration ratio up to 2.6, because a replicate averaged 40 splits where the estimate averages 300, so the split-to-split noise of the naming, largest at small budgets, entered every replicate; that is the deviation of Appendix S. With replicates of 300 splits, the interval the paper now prints, the ratio at eight clusters in the real world runs 0.86 to 1.31 by configuration and is 1.09 for the primary mean, and in the exchangeable null 1.13 to 1.45 at half and 1.32 to 1.68 at eight. The exchangeable null has no systematic disagreement, its two rules selecting different evaluators on half to three quarters of the draws with neither preferred in truth, so the *heterogeneous null* gives two evaluators the same within-cluster part and different between-cluster parts, so that their shares differ, pooling prefers the first, and every payoff difference between the two is zero in truth, the rest of the panel, 9 to 20 evaluators in all and a smaller naming problem than the 174-model panel graded, the noisier fillers of the cross world; its estimand is zero when the two rules select a filler equally often and departs from it by the fillers' shortfall times the difference of the two naming rates, so its truth is taken by Monte Carlo, 200 corpora of 60 splits. Run on 24 configurations, the 16 of the other synthetic worlds and the eight of the two validation sets, the two rules disagree on 16% to 55% of draws at half, where the truth lies within a point of zero on 22 configurations and at -3.1 and -2.8 on SummEval consistency and fluency, where the grouped mean already selects a filler more often than pooling does; the interval covers it on 99% of replications (96% at the lowest configuration) and lies wholly above zero, the reading that would support H1 with nothing between the two, on 0.5% (at most 2.0%, on 2 configurations); coverage here is read against the Monte Carlo truth of the world and the one-sided shares against a truth of exactly zero, which is what the world's construction gives when neither rule prefers a filler; the mean over the five RewardBench 2 configurations is covered on 100% and excludes zero on either side on 0.0% of replications, the WMT23 mean on 98% and 3.0%, and means over random subsets of five and three configurations exclude zero on 2.2% and 2.5%. At eight clusters the departure is general: the grouped mean, whose eight per-cluster coefficients separate the two from the fillers less surely than the pooled coefficient, which also reads the between-cluster part the fillers share with the first evaluator, selects a filler more often than pooling does, the only source of a negative truth here since the two structured evaluators pay alike and every filler less; the truth runs from -4.1 to $+0.0$ points, -4.1 to -3.3 on WMT23 and -3.4 on RewardBench 2 safety, and the interval, 0.96 to 1.60 times too wide, covers it on 97% of replications (94% at the lowest) and lies wholly above zero on 0.2%, so a reading at eight clusters that crosses zero says little either way, and one below zero may be the naming at that budget and not the evaluators. The percentile interval covers on 97% in the real world and 100% in the null world at half. These are the levels at which the best-of-n interval of the grouped mean against pooling is to be read in Tables 5 and 43, under the resampling model and conditional on the benchmarks. The other contrasts of those tables were run through the same pipeline in the real world alone (Table 74): the pairwise accuracy against pooling under best of n, the contrast whose main-text interval clears zero, is covered on 92% for the primary mean at half (bootstrap standard deviation 0.95 times the true one) and 97% at eight, per configuration on 95% at half and short on the same configurations, 81% on SummEval coherence, 86% on HANNA relevance and 86% on FRANK, the RewardBench 2 mean on 96% and the WMT23 mean on 96%; the grouped mean against pooling under the pairwise choice on 92% and 96% for the primary mean, 96% for the RewardBench 2 mean, the level of H1's pairwise grade, and 94% per configuration at half (82% at the lowest); and best of n on the labelled clusters against the grouped mean under best of n, H4's utility contrast, on 98% and 98% for the primary mean and 94% for the WMT23 mean at eight. Not run through the calibration: the grouped mean against a benchmark's own statistic, and the PPE, PRM800K, family-panel, candidate-count, leave-one-out and error intervals. What none of this measures is the uncertainty of generalising to a benchmark not in the six, which no resample of these can.

The worlds are built on the primary configurations, the real world and the heterogeneous null also on the two validation sets. The intervals reported for PRM800K, PPE, WMT24 and WMT25 come from bootstraps of the same kinds with no worlds of their own, and the fifth protocol's also resamples the measurement set, a scheme no world here runs.

AB Evaluator Rankings under the Two Components

Ranking the evaluators of a benchmark twice, by the protocol's own coefficient and by the within-cluster one, gives orderings that agree at a Kendall correlation of 0.28 on FRANK, 0.63 on WMT22 en-de and 0.58 on SummEval coherence. On FRANK, QAGS reaches 0.578 pooled with 0.065 within systems and FactCC 0.600 with 0.204; that benchmark already partials system identity out, and the decomposition recovers the choice from the labels rather than by convention. FRANK releases 15 metric score files, of which two, an entailment metric and a question-answering metric, do not score every summary; the analysis keeps the 13 that do, a set that differs from the decision experiment's by a second rule stated in Appendix H. The paired cluster bootstrap marks how many pairs change places with both intervals excluding zero: 14 of the 78 pairs among those metrics, 2 on WMT22 en-de between the segment-level and per-segment orderings, and none on SummEval. Counts on benchmarks with fewer than twenty clusters, which includes all three, are descriptive rather than tested at the nominal level, for the coverage reason of Appendix D; the level a paired interval has at nine to twenty clusters under the null is the one Appendix Z reports, 11.3% to 5.7% for r_T. The comparison resamples both factors of a crossed design, since which systems entered the comparison is part of the uncertainty of a statement about evaluator rankings.

AC Speech Assessment, MOS Prediction, and Cluster-Conditioned Evaluators

Identity-only predictors and partial-input baselines are established in single domains: speech quality and machine translation (Chinen et al., 2022; Perrella et al., 2024b), annotation artefacts and shortcut learning (Gururangan et al., 2018; Geirhos et al., 2020), and summary coherence, natural language inference and LLM-judge benchmarks (Steen & Markert, 2022; Poliak et al., 2018; Zheng et al., 2025). Cumlin (2026) and Pieper & Voran (2026) condition on the item for a rater-noise ceiling in MOS prediction. Sentence-level and speaker-level correlations were reported separately in early pronunciation scoring (Neumeyer et al., 1996; Franco et al., 1997), while speechocean762 (Zhang et al., 2021) is now scored by pooled phone-, word- and utterance-level correlations, with no speaker-level one (Gong et al., 2022; Chao & Chen, 2025; Yan et al., 2025a). MOS prediction keeps utterance-level and system-level correlations apart (Huang et al., 2022a), models listener effects (Leng et al., 2021; Huang et al., 2022b) and studies generalization across corpora (Cooper et al., 2022), and has standard-error analyses for listening tests (Camp et al., 2023) as language modelling has error bars and power analyses for benchmark means (Card et al., 2020; Miller, 2024). Evaluators that receive cluster information are not new either: hierarchical MOS models (Yoshimura et al., 2016), system and rater identity as inputs (Chinen et al., 2022), retrieval of related translations (Züfle et al., 2025), and cluster-mean conditioning in panel econometrics (Mundlak, 1978; Bell & Jones, 2015; Clarke & Polselli, 2026). We found no report of whether such conditioning improves the within-cluster component, which Appendix AD answers.

AD Cluster Information Moves Pooled Scores

Equation 1 implies that an evaluator can raise its pooled correlation without judging any item better, by weighting its between-cluster component more heavily. Two uses of cluster information test this, both available whenever several items of a cluster are scored together.

History pooling. For any evaluator output p, let $\bar{p}_c$ be its mean over the items of cluster c in the evaluation set. History pooling replaces p_i by

$$p'_i \;=\; a + b\,\bar{p}_{c(i)} + w\,\big(p_i - \bar{p}_{c(i)}\big), \tag{4}$$

with the three coefficients fit by least squares on held-out clusters, in two folds over clusters averaged over 20 random fold assignments, so that no item's label calibrates its own score. Every item of a

cluster then receives the same coefficients, so with $w > 0$ the order of items inside the cluster is unchanged and the pooled r_W moves only because the fitted slope re-weights clusters, by at most 0.013 on speechocean762. The gain in r_T comes from the between-cluster term. On speechocean762, history pooling raises the pooled correlation of the released GOPT checkpoint from 0.744 to 0.794, a paired gain of 0.051 with a 95% interval from 0.032 to 0.072, and the pooled correlation of the WavLM regression by 0.013 to 0.017 across the four aspects.

The size of the gain follows the decomposition. It is largest where an evaluator's own cluster-level estimate is noisy, as on FRANK, where pooling lifts QAGS from 0.578 to 0.809 while its within-system correlation moves only from 0.065 to 0.047, and the largest gain there is 0.388 for BLEU; it vanishes where the evaluator already ranks clusters almost perfectly, as on BVCC, where UTMOS22 moves from 0.885 to 0.886. An evaluator can therefore gain a large amount of apparent validity from identity information, or none, without any change in what it judges about an item, and the gains move published rankings: on speechocean762 the reproduced HMamba rises from 0.824 to 0.841, crossing the best published pooled correlation of 0.830 (Yan et al., 2025a), and GOPT and MultiPA, which pooled correlation cannot separate at $p = 0.887$, exchange places. Fitting the coefficients on the evaluation set itself, which the ITU-T standard for evaluating objective quality models allows through its per-experiment mapping (ITU-T, 2020), would raise them further, as calibration fitted on the labels raises an LLM judge's global correlation (Landesberg, 2026).

The second use of cluster information enters training instead: giving the evaluator the mean feature of its cluster raises within-speaker correlation on speechocean762 by 0.017 to 0.040 but does not transfer to BVCC or SOMOS, so it is a benchmark-specific baseline and not an architecture (Appendix AE).

AE Cluster-Mean Conditioning in Detail

Three forms receive identical inputs: concatenation $f([\phi(x_i), \mu_{c(i)}])$, the two-component form $N(\mu_{c(i)}) + W(\phi(x_i) - \mu_{c(i)})$ in which W is fit on within-cluster centred features and targets, and the plain regression $f(\phi(x_i))$ as reference. Each is a ridge regression on WavLM features or a one-hidden-layer network on their 256 principal components, with hyperparameters selected by cross-validation over training speakers only. Table 69 reports all three on speechocean762. Concatenation gains 0.018 to 0.031 in r_T beside its within-speaker gains, and the two-component form differs from concatenation by -0.013 to 0.004 in r_W and -0.012 to -0.001 in r_T; its paired interval excludes zero, read at the level such an interval has at 125 clusters (Appendix Z), in r_T for all four aspects with the network, for accuracy with ridge, and in r_W for the total score with ridge. On BVCC the three forms reach utterance-level Spearman correlations of 0.877, 0.875 and 0.869, and on SOMOS 0.703, 0.695 and 0.699. The within-speaker gains on speechocean762 remain small next to the distance between any evaluator and what rater agreement allows.

Table 47: The WMT24 configurations of the third protocol, as the frozen adapter built them: systems kept (human references and the outlier system of mt-metrics-eval left out), source segments rated for every kept system, documents, metrics of the graded panel (the year's primaries under the standard reference or the source that score every item, the three sentinels left out), the mean number of segments in a selection half, and per configuration the grouped mean's ordering advantage over pooling and the utility gain of the evaluator it selects at half, in points.

							best of n		pairwise choice	
pair	label	systems	segments	documents	metrics	half	advantage	gain	advantage	gain
en-de	MQM	16	487 of 998	171	23	241	+17.8	+1.5	+18.0	+0.8
en-es	MQM	13	623 of 998	147	23	306	+7.0	+1.4	+14.9	+0.0
ja-zh	MQM	13	560 of 722	197	23	281	+7.4	−11.3	+14.9	−1.5
cs-uk	ESA	11	1955 of 2317	303	20	974	+3.3	+1.2	+5.5	+0.6
en-cs	ESA	15	572 of 998	149	20	283	+15.2	+1.6	+16.6	+0.5
en-hi	ESA	10	635 of 998	171	20	315	+16.3	+4.5	+15.7	+0.3
en-is	ESA	10	635 of 998	171	20	319	+13.3	−0.9	+14.4	−0.3
en-ja	ESA	12	635 of 998	171	20	317	+15.9	+1.2	+14.0	+0.3
en-ru	ESA	13	635 of 998	171	20	316	+13.5	+1.3	+12.0	+0.3
en-uk	ESA	10	635 of 998	171	20	318	+6.3	+5.0	+8.6	+1.8
en-zh	ESA	12	635 of 998	171	20	315	+7.0	+2.0	+12.6	+1.0

Table 48: The third protocol, frozen and time-stamped before any WMT24 metric score was read: set means at half the segments of the grouped mean's ordering advantage over pooling (points of selection error over evaluator pairs) and of the utility gain of the evaluator it selects (points), with their grades. The line of M2 is the least-squares line of gain on advantage over the 21 earlier configurations outside machine translation, fixed in the protocol. Intervals: the estimate ± 1.96 standard deviations of the corpus bootstrap of the set mean; supp. supported, cons. consistent (the predicted side, the interval covering the threshold), not supp. not supported, contr. contradicted. The MQM set is the protocol's primary test and the ESA set its secondary one; the three sentinel metrics of the year are left out of both panels (Table 49 keeps them).

	WMT24 MQM (3)	WMT24 ESA (8)
best of n		
M1 advantage at half ≥ 8	+10.8 [+6.7, +14.8] cons.	+11.3 [+9.5, +13.2] supp.
M2 gain at half below the line	−2.8 [−9.2, +3.7] supp. (line +6.3)	+2.0 [−0.3, +4.3] supp. (line +6.6)
M3 gain at half below 3	cons.	cons.
M4 gain at 32 less gain at half > 0	+3.5 [−2.5, +9.4] cons.	−0.4 [−2.4, +1.6] not supp.
M4 advantage at half less at 32 > 0	+3.8 [+2.0, +5.6] supp.	+5.8 [+4.5, +7.1] supp.
pairwise choice		
M1 advantage at half ≥ 8	+15.9 [+11.9, +19.9] supp.	+12.4 [+10.1, +14.8] supp.
M2 gain at half below the line	−0.2 [−1.1, +0.7] supp. (line +1.7)	+0.6 [+0.1, +1.1] supp. (line +1.5)
M3 gain at half below 1	supp.	cons.

Table 49: The third protocol on all four sets: as Table 48, with the panels that keep the year's three sentinel metrics, two of which are constant inside a source segment by construction.

	WMT24 MQM (3)	WMT24 ESA (8)	MQM, with sentinels (3)	ESA, with sentinels (8)
best of n				
M1 advantage at half ≥ 8	+10.8 [+6.7, +14.8] cons.	+11.3 [+9.5, +13.2] supp.	+13.2 [+9.2, +17.2] supp.	+8.9 [+7.2, +10.6] cons.
M2 gain at half below the line	−2.8 [−9.2, +3.7] supp. (line +6.3)	+2.0 [−0.3, +4.3] supp. (line +6.6)	−2.8 [−8.7, +3.2] supp. (line +7.7)	+2.0 [−0.2, +4.2] supp. (line +5.3)
M3 gain at half below 3	cons.	cons.	cons.	cons.
M4 gain at 32 less gain at half > 0	+3.5 [−2.5, +9.4] cons.	−0.4 [−2.4, +1.6] not supp.	+3.7 [−1.9, +9.2] cons.	−0.6 [−2.6, +1.5] not supp.
M4 advantage at half less at 32 > 0	+3.8 [+2.0, +5.6] supp.	+5.8 [+4.5, +7.1] supp.	+3.2 [+1.3, +5.2] supp.	+3.9 [+2.9, +4.9] supp.
pairwise choice				
M1 advantage at half ≥ 8	+15.9 [+11.9, +19.9] supp.	+12.4 [+10.1, +14.8] supp.	+18.4 [+14.7, +22.1] supp.	+11.2 [+9.3, +13.0] supp.
M2 gain at half below the line	−0.2 [−1.1, +0.7] supp. (line +1.7)	+0.6 [+0.1, +1.1] supp. (line +1.5)	−0.2 [−1.1, +0.7] supp. (line +1.8)	+0.6 [+0.1, +1.1] supp. (line +1.4)
M3 gain at half below 1	supp.	cons.	supp.	cons.

Table 50: The metric each rule names reading every segment of a WMT24 configuration (graded panels, sentinels left out), with its rank among the panel's K metrics by best-of-n recovery on those segments; last column, the share of the 300 selection halves on which pooling and the grouped mean name the every-segment best, in percent. Read after the grading (`scripts/decision/wmt24_winners_r16.py`, a replay of the frozen analysis that reproduces its released results exactly).

pair	label	K	r_T	r_W	grp	pairwise	own	best named, %
en-de	MQM	23	XCOMET (2)	XCOMET (2)	MM-hybrid (1)	MM-hybrid (1)	MM-hybrid (1)	16 / 100
en-es	MQM	23	MM-hybrid (3)	MM-hybrid (3)	COMET-22 (1)	COMET-22 (1)	Bright-QE (18)	0 / 58
ja-zh	MQM	23	MM-hybrid (1)	MM-hybrid (1)	GEMBA-ESA (7)	MetricX-24 (5)	MM-hybrid (1)	97 / 3
cs-uk	ESA	20	COMET-22 (3)	Prism-M (2)	MM-hybrid (1)	MM-hybrid (1)	MM-hybrid (1)	0 / 84
en-cs	ESA	20	MetricX-24 (2)	MetricX-24 (2)	MetricX-24 (2)	MetricX-24 (2)	MetricX-24 (2)	0 / 2
en-hi	ESA	20	MetricX-24 (5)	MetricX-24 (5)	MetricX-24 (5)	MetricX-24 (5)	MetricX-24 (5)	0 / 37
en-is	ESA	20	MetricX-24 (1)	MetricX-24 (1)	MetricX-24 (1)	MetricX-24 (1)	MM-hybrid (2)	89 / 66
en-ja	ESA	20	MetricX-24 (5)	MetricX-24 (5)	MetricX-24 (5)	MetricX-24 (5)	MetricX-24 (5)	0 / 1
en-ru	ESA	20	MetricX-24 (4)	MetricX-24 (4)	MetricX-24 (4)	MetricX-24 (4)	MM-hybrid (2)	0 / 0
en-uk	ESA	20	GEMBA-ESA (11)	XCOMET (2)	XCOMET (2)	XCOMET (2)	XCOMET (2)	2 / 8
en-zh	ESA	20	GEMBA-ESA (7)	COMET-22 (1)	MM-hybrid (2)	MM-hybrid (2)	MetricX-24 (6)	8 / 28

Table 51: The WMT24 reading on other panels, not graded and read after the third protocol was graded: set means at half the segments of the grouped mean's ordering advantage over pooling and of the utility gain of its evaluator, in points, ± 1.96 standard deviations of a corpus bootstrap of 50 corpora of 20 splits (smaller than the graded runs'). The en-es row holds the same systems and segments under the MQM label and under the ESA label, one configuration each (`scripts/decision/wmt24_panels_r16.py`, a rerun of the frozen analysis whose unchanged panel reproduces the graded results exactly).

	best of n		pairwise choice	
panel	advantage	gain	advantage	gain
MQM (3), graded panel	+10.8 [+6.9, +14.6]	−2.8 [−10.2, +4.7]	+15.9 [+11.8, +20.0]	−0.2 [−1.1, +0.7]
MQM, without GEMBA-ESA	+11.9 [+7.7, +16.1]	−0.3 [−7.0, +6.4]	+17.3 [+12.6, +22.0]	+0.4 [−0.4, +1.1]
MQM, reference-free metrics only	+5.8 [−1.5, +13.1]	−2.1 [−8.2, +4.0]	+16.3 [+10.4, +22.3]	+0.4 [−0.9, +1.8]
ESA (8), without GEMBA-ESA	+11.2 [+9.2, +13.2]	+1.6 [−0.7, +3.9]	+12.2 [+9.5, +14.9]	+0.4 [+0.0, +0.9]
ESA, reference-free metrics only	+15.0 [+11.1, +18.9]	+4.0 [+1.7, +6.3]	+17.6 [+12.6, +22.5]	+1.1 [+0.5, +1.6]
en-es, items rated under both labels	+8.9 [+2.8, +15.0]	+5.6 [−5.9, +17.2]	+7.3 [+0.5, +14.1]	+0.2 [−1.4, +1.9]

Table 52: The fourth protocol's frozen run beside the reruns outside it, read after the grading: every rated system kept (18 to 20 a language pair), the merged ESA score as the label, every system but those built on a judge's model, and the LLM-judge panel without Gemini 2.5 Pro, each graded by the same graders (`scripts/decision/run_wmt25_variants.py`); the headline quantities, with every N4, M2 and M3 grade. Best of n unless stated, points, set means with ± 1.96 corpus-bootstrap standard deviations; grades as in Table 53. The merged label leaves the MQM set unchanged.

	frozen, primaries	all systems, primaries	merged ESA label, primaries	frozen, LLM judges	all systems, LLM judges	merged ESA label, LLM judges	all systems but the judges' own, LLM judges	LLM judges without Gemini 2.5 Pro
advantage at half, MQM (N1)	+9.4 supp.	+10.3 supp.	+9.4 supp.	+6.4 supp.	+7.8 supp.	+6.4 supp.	+7.3 supp.	+7.8 supp.
advantage at half, ESA (N1)	+9.4 supp.	+11.8 supp.	+10.8 supp.	+3.5 supp.	+2.7 supp.	+4.8 supp.	+4.4 supp.	+3.9 supp.
gain at half, MQM (M2)	+4.7 cons.	+5.1 cons.	+4.7 cons.	+0.5 cons.	−1.2 cons.	+0.5 cons.	+0.9 cons.	+4.1 cons.
gain at half, ESA (M2)	+1.5 supp.	+3.1 supp.	+1.0 supp.	+1.1 cons.	+3.4 contr.	+1.2 supp.	+0.9 supp.	+0.9 cons.
label of the third protocol's rule, ESA	as the grades stand	confirmed	confirmed	failed	failed	failed	failed	failed
r at 8 segments; at half	0.58; 0.56	0.69; 0.57	0.47; 0.21	0.44; 0.04	0.21; −0.28	0.28; 0.21	0.56; 0.04	0.61; 0.24
N2a; N2b	supp.; cons.	cons.; cons.	cons.; cons.	cons.; cons.	cons.; cons.	cons.; cons.	cons.; cons.	cons.; cons.
gain at 8, grp, MQM	+2.0 [−5.3, +9.2]	+7.0 [+0.9, +13.1]	+2.0 [−5.3, +9.2]	+6.4 [+1.0, +11.7]	+8.3 [+2.3, +14.4]	+6.4 [+1.0, +11.7]	+6.9 [+1.1, +12.6]	+7.3 [+1.5, +13.1]
gain at 8, grp, ESA	−0.4 [−2.1, +1.4]	+2.8 [+0.5, +5.1]	−0.4 [−2.1, +1.3]	−0.1 [−1.4, +1.3]	+2.2 [+1.0, +3.3]	+0.2 [−1.1, +1.6]	+1.0 [−0.2, +2.2]	+0.4 [−1.0, +1.8]
gain at 8, acc^*_{eq}, MQM	−3.9 [−10.3, +2.5]	−0.3 [−6.0, +5.5]	−3.9 [−10.3, +2.5]	+7.0 [+1.0, +13.1]	+8.6 [+2.2, +15.0]	+7.0 [+1.0, +13.1]	+8.9 [+2.8, +15.1]	+6.5 [+0.4, +12.6]
gain at 8, acc^*_{eq}, ESA	+0.2 [−1.8, +2.2]	+2.0 [−0.5, +4.5]	−0.2 [−2.1, +1.8]	+5.6 [+3.5, +7.7]	+6.4 [+4.9, +7.9]	+5.5 [+3.7, +7.4]	+6.4 [+4.7, +8.1]	+4.6 [+2.8, +6.3]
N4 best of n on labels vs r_T at 8, MQM	−0.0 cons.	+0.4 not supp.	−0.0 cons.	+4.3 not supp.	+3.8 not supp.	+4.3 not supp.	+5.8 not supp.	+6.0 contr.
N4 best of n on labels vs r_T at 8, ESA	−1.5 cons.	+1.3 not supp.	−1.5 cons.	+0.5 not supp.	+1.5 contr.	+0.9 not supp.	+1.3 not supp.	+0.7 not supp.
M3 gain at half below 3, MQM; ESA	not supp.; cons.	not supp.; not supp.	not supp.; cons.	cons.; supp.	cons.; not supp.	cons.; supp.	cons.; supp.	not supp.; supp.
M2 pairwise choice, MQM; ESA	cons.; supp.	cons.; supp.	cons.; supp.	cons.; supp.	cons.; supp.	cons.; supp.	not supp.; supp.	not supp.; supp.

Table 53: The fourth protocol's grades, every set and both panels: WMT25's two MQM language pairs and 13 ESA pairs (en-mas_KE left out, Deviation 1), scored by the year's primary metrics and by thirteen general-purpose LLMs prompted as judges. Set means in points at the budget named, ± 1.96 standard deviations of the frozen corpus bootstrap (200 corpora of 40 splits); M2's line value in parentheses; N2a and N2b, correlations over the configurations of both sets, with the 2.5th and 97.5th percentiles of 10,000 hierarchical resamples (the MQM columns hold N2a, the ESA columns N2b); N5, the difference of the two sets. M1 to M4 are the third protocol's hypotheses graded by its frozen grader, M4's interval unpaired on the MQM sets (Deviation 2); an M1 estimate printed $+8.0$ and graded not supported is 7.99. Refitted without HANNA relevance the best-of-n line would give about 3.1 at the primaries' advantages (Appendix U). supp. supported, cons. consistent, not supp. not supported, contr. contradicted.

	MQM (2)		ESA (13)	
	estimate	grade	estimate	grade
primaries				
best of n				
N1 advantage at half > 0	+9.4 [+3.8, +15.0]	supp.	+9.4 [+7.0, +11.7]	supp.
M1 advantage at half ≥ 8	+9.4 [+3.8, +15.0]	cons.	+9.4 [+7.0, +11.7]	cons.
M2 gain at half below the line	+4.7 [−4.0, +13.4] (+5.6)	cons.	+1.5 [−0.8, +3.9] (+5.5)	supp.
M3 gain at half below 3	+4.7 [−4.0, +13.4]	not supp.	+1.5 [−0.8, +3.9]	cons.
M4 gain falls, 32 to half	−0.2 [−13.2, +12.7]	not supp.	−1.0 [−2.6, +0.6]	not supp.
M4 advantage rises	+1.0 [−6.8, +8.9]	cons.	+2.7 [+1.7, +3.6]	supp.
N4 best of n on 8 labelled segments vs $r_T < 0$	−0.0 [−7.8, +7.8]	cons.	−1.5 [−3.7, +0.7]	cons.
N2a $r_8 > 0$; N2b $r_8 - r_{C/2} > 0$ (all 15)	+0.58 [+0.03, +0.82]	supp.	+0.02 [−0.48, +0.54]	cons.
N5 ESA minus MQM gain at half > 0	−3.2 [−12.2, +5.9]	not supp.		
pairwise choice				
N1 advantage at half > 0	+6.3 [+0.5, +12.2]	supp.	+9.1 [+6.6, +11.6]	supp.
M1 advantage at half ≥ 8	+6.3 [+0.5, +12.2]	not supp.	+9.1 [+6.6, +11.6]	cons.
M2 gain at half below the line	−0.2 [−2.2, +1.9] (+1.1)	cons.	+0.4 [−0.2, +0.9] (+1.3)	supp.
M3 gain at half below 1	−0.2 [−2.2, +1.9]	cons.	+0.4 [−0.2, +0.9]	supp.
N2a $r_8 > 0$; N2b $r_8 - r_{C/2} > 0$ (all 15)	+0.71 [+0.08, +0.83]	supp.	+0.13 [−0.49, +0.55]	cons.
N5 ESA minus MQM gain at half > 0	+0.6 [−1.5, +2.7]	cons.		
r at 8 and at half within ESA, best of n	0.56 and 0.82			
r at 8 and at half within ESA, pairwise choice	0.71 and 0.78			
label of the third protocol's rule	MQM: as the grades stand		ESA: as the grades stand	
labels of the ordering and gain clauses	confirmed; as the grades stand			
LLM judges				
best of n				
N1 advantage at half > 0	+6.4 [+1.5, +11.3]	supp.	+3.5 [+2.1, +4.9]	supp.
M1 advantage at half ≥ 8	+6.4 [+1.5, +11.3]	not supp.	+3.5 [+2.1, +4.9]	contr.
M2 gain at half below the line	+0.5 [−8.1, +9.1] (+3.9)	cons.	+1.1 [−0.5, +2.7] (+2.3)	cons.
M3 gain at half below 3	+0.5 [−8.1, +9.1]	cons.	+1.1 [−0.5, +2.7]	supp.
M4 gain falls, 32 to half	+3.8 [−8.4, +16.0]	cons.	−0.1 [−1.2, +1.0]	not supp.
M4 advantage rises	−1.7 [−8.6, +5.2]	not supp.	−0.2 [−0.9, +0.5]	not supp.
N4 best of n on 8 labelled segments vs $r_T < 0$	+4.3 [−1.3, +10.0]	not supp.	+0.5 [−1.1, +2.1]	not supp.
N2a $r_8 > 0$; N2b $r_8 - r_{C/2} > 0$ (all 15)	+0.44 [−0.08, +0.78]	cons.	+0.40 [−0.44, +0.76]	cons.
N5 ESA minus MQM gain at half > 0	+0.6 [−8.1, +9.4]	cons.		
pairwise choice				
N1 advantage at half > 0	+8.0 [+3.0, +13.0]	supp.	+3.8 [+2.3, +5.3]	supp.
M1 advantage at half ≥ 8	+8.0 [+3.0, +13.0]	not supp.	+3.8 [+2.3, +5.3]	contr.
M2 gain at half below the line	+0.4 [−1.9, +2.8] (+1.2)	cons.	+0.3 [−0.2, +0.8] (+0.9)	supp.
M3 gain at half below 1	+0.4 [−1.9, +2.8]	cons.	+0.3 [−0.2, +0.8]	supp.
N2a $r_8 > 0$; N2b $r_8 - r_{C/2} > 0$ (all 15)	+0.48 [−0.10, +0.80]	cons.	+0.66 [−0.34, +0.82]	cons.
N5 ESA minus MQM gain at half > 0	−0.2 [−2.6, +2.3]	not supp.		
r at 8 and at half within ESA, best of n	0.22 and 0.28			
r at 8 and at half within ESA, pairwise choice	0.08 and −0.19			
label of the third protocol's rule	MQM: failed		ESA: failed	
labels of the ordering and gain clauses	confirmed; as the grades stand			

Table 54: The WMT25 configurations of the fourth protocol, as the frozen adapters built them: systems kept of the machine-translation systems rated (every system is rated on the same segments, so the shortfall rule gained no segment and kept the eight last by name), source segments (usually paragraphs) rated for every kept system, documents, metrics of each panel that score every item, the mean number of segments in a selection half, and under best of n the grouped mean's ordering advantage over pooling (adv.) and the utility gain of its evaluator at 8 segments and at half, in points, for each panel.

					metrics			primaries				LLM judges			
pair	label	systems	segments	docs	prim.	LLM	half	adv. 8	gain 8	adv. $C/2$	gain $C/2$	adv. 8	gain 8	adv. $C/2$	gain $C/2$
cs-de_DE	ESA	8 of 20	231 of 457	128	11	13	116	+5.0	+1.4	+20.4	+8.1	−1.7	−0.8	+1.0	+0.2
cs-uk_UA	ESA	8 of 18	225 of 405	131	11	13	112	+1.8	−6.1	+11.4	+0.1	+1.2	+1.5	+4.3	+0.7
en-ar_EG	ESA	8 of 18	199 of 332	51	11	13	99	+6.3	+0.4	+10.4	+0.0	+3.2	−0.9	+1.4	−0.4
en-bho_IN	ESA	8 of 18	189 of 332	50	11	13	93	−0.5	+1.0	−0.7	−0.8	+1.8	−0.1	+1.0	+0.0
en-cs_CZ	ESA	8 of 19	212 of 384	60	12	13	107	+7.6	+0.6	+22.0	+6.8	+5.0	−1.1	+2.9	+0.7
en-et_EE	ESA	8 of 18	190 of 332	52	11	13	95	−0.1	−1.3	−1.0	−5.3	+2.4	−5.2	−0.6	−0.1
en-is_IS	ESA	8 of 18	202 of 332	53	11	13	95	+2.3	+0.6	+5.0	+1.5	+5.4	−3.6	+6.5	−0.6
en-it_IT	ESA	8 of 18	215 of 332	54	11	13	108	−4.8	−2.9	−5.1	−0.6	+7.5	+1.6	+9.5	+0.8
en-ja_JP	ESA	8 of 18	189 of 332	51	12	13	96	+0.6	−0.1	+3.7	−0.8	+4.2	+1.3	+2.7	+3.4
en-mas_KE	ESA	8 of 18	185 of 332	49	0	0	left out: no metric of either panel scores every item (Deviation 1)								
en-ru_RU	ESA	8 of 18	200 of 332	52	11	13	100	+4.8	−1.0	+13.9	+1.8	+2.9	+0.4	+3.5	+3.7
en-sr_Cyrl_RS	ESA	8 of 18	206 of 332	50	11	13	102	+7.5	+6.0	+18.7	+8.4	+5.4	+4.1	+6.3	+5.7
en-uk_UA	ESA	8 of 18	199 of 332	52	11	13	99	+1.8	−1.5	+14.8	+2.6	+3.5	+0.4	+3.0	−0.4
en-zh_CN	ESA	8 of 18	198 of 332	51	12	13	99	+0.5	−2.1	+8.2	−2.2	+4.5	+1.7	+4.3	+0.7
en-ko_KR	MQM	8 of 19	112 of 332	74	11	13	56	+5.4	+4.0	+7.8	+13.4	+10.5	+4.2	+8.2	−2.6
ja-zh_CN	MQM	8 of 18	158 of 334	74	11	13	79	+4.6	−0.1	+11.0	−4.0	+5.9	+8.5	+4.6	+3.6

Table 55: The regret each rule's pick leaves on the evaluation half of the WMT25 sets (points of the payoff; set means), at 8 and 32 labelled segments and at half; pairwise and best of n are the pairwise accuracy and best of n computed on the labelled segments, own is acc^*_{eq}; and at eight segments the number of labelled segments at which pooling's pick leaves what each statistic's pick leaves (interpolated in $\log_2$ of the budget on the grid; $<$8, pooling at eight already does at least as well; $>$grid, beyond the grid), and in the row of 32 segments the same at sixteen. Budgets larger than a configuration's smallest selection half rest on the draws whose half reaches them. Described by the protocol, not graded.

				regret of the pick						pooling's equivalent at 8			
panel	set	payoff	b	r_T	r_W	grp	pairwise	own	best of n	r_W	grp	pairwise	own
primaries	MQM	best of n	8	18.1	16.5	16.2	15.8	22.0	18.2	11.5	16.6	18.4	$<$8
			32	14.3	12.7	9.8	9.9	19.8	12.9	at 16: 28.3	51.8	51.3	5.7
			$C/2$	13.2	13.0	8.6	8.1	18.2	9.2				
		pairwise choice	8	3.8	3.4	3.7	3.5	4.5	4.5	10.7	8.7	9.6	$<$8
			32	2.3	2.1	2.2	2.1	3.6	3.2	at 16: 19.7	18.0	18.9	8.1
			$C/2$	1.6	1.8	1.8	1.4	3.0	2.4				
	ESA	best of n	8	12.7	12.7	13.0	12.6	12.5	14.1	$<$8	$<$8	8.2	8.9
			32	9.6	9.6	9.1	8.7	8.6	10.5	at 16: 14.6	16.4	20.8	23.2
			$C/2$	8.3	7.5	6.8	6.6	6.8	8.8				
		pairwise choice	8	2.7	2.7	2.9	2.8	3.0	3.5	$<$8	$<$8	$<$8	$<$8
			32	1.9	1.8	1.8	1.7	1.7	2.5	at 16: 15.6	16.1	19.7	17.3
			$C/2$	1.6	1.3	1.2	1.1	1.1	2.0				
LLM judges	MQM	best of n	8	17.9	13.8	11.5	9.5	10.8	13.5	24.8	33.0	43.9	36.3
			32	11.7	9.6	7.4	5.6	6.8	11.3	at 16: 32.7	46.8	$>$grid	$>$grid
			$C/2$	6.3	7.0	5.8	4.8	6.0	10.9				
		pairwise choice	8	5.9	4.4	3.4	2.6	3.4	4.3	20.5	29.9	41.2	29.7
			32	3.2	2.6	1.7	1.0	2.0	2.5	at 16: 31.2	43.0	$>$grid	42.8
			$C/2$	1.4	1.6	1.0	0.7	1.6	1.8				
	ESA	best of n	8	9.1	10.0	9.2	6.8	3.5	8.6	$<$8	$<$8	19.9	$>$grid
			32	5.8	5.4	4.8	3.4	2.8	4.7	at 16: 17.2	21.7	61.0	$>$grid
			$C/2$	4.2	3.1	3.1	2.6	2.3	3.0				
		pairwise choice	8	2.6	2.8	2.6	1.9	1.0	2.6	$<$8	$<$8	17.2	70.1
			32	1.5	1.3	1.1	0.7	0.5	1.3	at 16: 18.0	21.7	51.8	$>$grid
			$C/2$	0.8	0.6	0.5	0.3	0.3	0.8				

Table 56: How the WMT25 evaluators score the kept items, read after the grading (`scripts/decision/wmt25_described_r17.py`): the median over language pairs of the number of distinct scores an evaluator gives a pair's items, and the mean share of segments in which two or more of the eight systems share its top score, so that best of n through it is a draw. The two ESA ratings' agreement within a segment, by language pair: en-ja 0.04, en-zh 0.10, en-ru 0.21, en-it 0.24, cs-de 0.30, en-cs 0.30, cs-uk 0.32, en-et 0.33, en-uk 0.34, en-sr 0.39, en-is 0.39, en-bho 0.70, en-ar 0.84.

panel	evaluator	distinct scores	tie at the top, %
LLM judges	Llama-4-Maverick	8	98
	AyaExpanse-8B	8	88
	Claude-4	10	86
	AyaExpanse-32B	12	82
	Llama-3_1-8B	9	81
	DeepSeek-V3	10	80
	Mistral-7B	10	74
	CommandR7B	16	74
	CommandA	8	71
	Qwen3-235B	13	69
	Qwen2_5-7B	17	66
	GPT-4_1	26	54
	Gemini2_5-Pro	69	38
primary metrics	Q_Relative-MQM	84	74
	mr7_2_1	61	37
	comet-multi	43	27
	TASER-No-Ref	51	25
	GEMBA-v2	1120	9
	EnsembleSlick	616	9
	SEGALE-QE	1384	1
	UvA-MT	1583	1
	rankedCOMET	1591	0
	MetricX-25	1591	0
	hw-tsc	1590	0
	Polycand-2	1590	0

Table 57: The fifth protocol on PPE: every rule read off the labelled prompts of the pool, the judge it names paid its best-of-16 accuracy on the 192 fixed prompts a set (points, the five sets' mean). Gain: over pooling's pick; advantage: pooling's error over judge pairs minus the rule's; intervals, the estimate ± 1.96 standard deviations of 200 replicates resampling the pool's and the measurement set's prompts (the protocol's). Hit: the share of draws naming the best judge on the measurement set. Protocol and grades: Appendix V and Table 42. Half the pool is 48 prompts, 43 on panel B's MATH, whose pool lost 10 (Deviation 1); panel B's measurement set drops the 70 responses a reward model has no score for, GPQA keeping 190 prompts; panel B's intervals at 24 prompts rest on 192 replicates, 8 of MATH being undefined. These intervals resample the measurement set as well as the pool and lie outside the calibration of Appendix AA; their percentile intervals, which change no grade, are in `results/ppe_llm_report.json`.

rule	accuracy	gain over r_T	hit, %	error, %	advantage over r_T
Panel A, the LLM judges (7 evaluators)					
8 prompts; best judge 61.8, mean judge 58.5					
r_T	59.0	–	18	43.6	–
r_W	59.0	+0.0 [−0.3, +0.4]	16	43.6	+0.0 [−2.6, +2.6]
grp	58.9	−0.1 [−0.6, +0.3]	14	43.6	−0.0 [−3.4, +3.3]
pairwise acc.	59.4	+0.4 [−0.1, +0.9]	18	40.4	+3.2 [−0.4, +6.7]
best of 16 on labels	59.5	+0.5 [−0.1, +1.1]	20	41.3	+2.3 [−2.1, +6.6]
acc^*_{eq}	59.3	+0.3 [−0.1, +0.7]	18	44.3	−0.7 [−4.0, +2.6]
24 prompts; best judge 61.8, mean judge 58.5					
r_T	59.4	–	23	39.7	–
r_W	59.5	+0.1 [−0.6, +0.7]	17	40.4	−0.7 [−5.7, +4.3]
grp	59.3	−0.1 [−1.0, +0.7]	15	39.0	+0.7 [−5.4, +6.7]
pairwise acc.	59.9	+0.5 [−0.4, +1.3]	19	35.4	+4.3 [−2.0, +10.6]
best of 16 on labels	59.9	+0.5 [−0.5, +1.4]	21	36.9	+2.8 [−5.0, +10.5]
acc^*_{eq}	59.6	+0.2 [−0.6, +1.0]	19	43.2	−3.6 [−9.9, +2.7]
half the pool (48); best judge 61.8, mean judge 58.5					
r_T	59.8	–	29	35.5	–
r_W	59.8	+0.0 [−0.7, +0.8]	20	37.4	−1.9 [−7.7, +3.9]
grp	59.6	−0.2 [−1.2, +0.9]	14	35.0	+0.4 [−6.4, +7.3]
pairwise acc.	60.1	+0.3 [−0.6, +1.3]	17	31.9	+3.6 [−3.5, +10.7]
best of 16 on labels	60.0	+0.3 [−0.8, +1.3]	23	33.3	+2.1 [−6.5, +10.7]
acc^*_{eq}	59.9	+0.1 [−0.8, +1.0]	21	43.2	−7.8 [−15.0, −0.5]
Panel B, the LLM judges with the reward models and length (13 evaluators)					
8 prompts; best evaluator 64.3, mean evaluator 54.8					
r_T	58.6	–	13	30.8	–
r_W	59.7	+1.1 [+0.4, +1.9]	23	27.9	+2.9 [+1.0, +4.7]
grp	59.9	+1.3 [+0.3, +2.3]	25	27.4	+3.3 [+1.2, +5.4]
pairwise acc.	60.1	+1.6 [+0.4, +2.7]	29	28.1	+2.7 [+0.5, +4.9]
best of 16 on labels	58.6	+0.0 [−1.2, +1.3]	25	32.3	−1.5 [−4.1, +1.1]
acc^*_{eq}	58.9	+0.4 [−0.4, +1.1]	20	34.9	−4.1 [−6.7, −1.6]
24 prompts; best evaluator 64.3, mean evaluator 54.8					
r_T	59.6	–	15	25.2	–
r_W	61.0	+1.4 [+0.1, +2.7]	31	21.9	+3.3 [+0.2, +6.4]
grp	61.0	+1.4 [−0.2, +2.9]	31	20.8	+4.4 [+1.2, +7.6]
pairwise acc.	61.5	+1.9 [+0.2, +3.6]	36	20.6	+4.6 [+1.1, +8.2]
best of 16 on labels	61.0	+1.3 [−0.4, +3.1]	32	24.7	+0.6 [−3.6, +4.8]
acc^*_{eq}	60.3	+0.7 [−0.6, +1.9]	29	31.3	−6.1 [−10.3, −1.9]
half the pool (48); best evaluator 64.3, mean evaluator 54.8					
r_T	60.0	–	17	22.6	–
r_W	61.9	+2.0 [+0.5, +3.4]	41	19.1	+3.5 [+0.2, +6.9]
grp	61.4	+1.4 [−0.3, +3.1]	32	17.8	+4.9 [+1.4, +8.3]
pairwise acc.	62.1	+2.1 [+0.3, +3.9]	37	16.6	+6.0 [+2.2, +9.8]
best of 16 on labels	61.8	+1.8 [−0.1, +3.7]	37	19.8	+2.8 [−1.8, +7.5]
acc^*_{eq}	61.1	+1.1 [−0.3, +2.5]	39	29.5	−6.9 [−11.5, −2.2]
after the grading, not graded					
mean of the seven LLM judges' ratings, no labels	65.3	–	–	–	–

Table 58: The seven LLM judges of the fifth protocol's panel on PPE's measurement set: best-of-16 accuracy of the judge's top response (a tie at the top counted as the mean label of the tied responses), by set and on the five sets' mean; the most distinct ratings used on one set; the share of response pairs inside a prompt given equal ratings; the share of prompts on which several responses share the top rating; and the share of responses parsed (the rest imputed at 5.5; for Qwen3-30B-A3B-2507 and Qwen3-Next-80B-A3B the values and ties count the imputed MBPP+ ratings). Judges dropped by the completion rule: Qwen3-8B, Qwen3-14B, Qwen3-235B-A22B, Qwen3-235B-A22B-2507, Kimi-K2.

judge	MMLU-Pro	MATH	GPQA	IFEval	MBPP+	mean	values	tied pairs, %	tied top, %	parsed, %
Qwen3-Next-80B-A3B	71.5	57.2	49.1	56.4	68.7	60.6	10	48	82	98.6
DeepSeek-V3.2	69.9	50.8	51.2	56.4	70.2	59.7	10	47	79	100.0
Qwen-Flash	69.7	59.7	46.5	52.2	67.9	59.2	10	47	79	100.0
Qwen3-30B-A3B-2507	68.8	59.8	46.1	52.4	65.5	58.5	10	49	79	95.0
DeepSeek-V3	69.7	48.9	48.6	53.5	66.9	57.5	9	52	86	100.0
Qwen3-32B	66.9	49.3	48.4	54.4	67.6	57.3	9	53	88	100.0
Qwen3.5-Flash	66.2	49.0	49.1	52.7	65.1	56.4	9	69	92	100.0

Table 59: The fifth protocol by set. Top: each kept judge's share of response pairs inside a prompt given equal ratings and share of prompts whose top rating several responses share, in %, on the measurement set (the imputed MBPP+ ratings of two judges included). Bottom, described after the grading and not graded: best-of-16 accuracy on the measurement set of the best single judge, of the mean and the median of the seven judges' ratings, which read no label, and of the best reward model of the paper's PPE panel, and pooling's regret with eight labelled prompts (the best judge of the set minus pooling's pick), in points. Distinct values by judge and set are in `results/ppe_llm_targets_A.json`.

	MMLU-Pro	MATH	GPQA	IFEval	MBPP+
Qwen3-Next-80B-A3B	40/82	36/74	33/68	60/92	68/95
DeepSeek-V3.2	40/81	43/79	44/65	48/78	62/91
Qwen-Flash	38/73	32/69	39/67	59/90	66/96
Qwen3-30B-A3B-2507	38/74	32/69	39/64	58/91	75/96
DeepSeek-V3	44/84	43/85	42/73	62/92	66/95
Qwen3-32B	45/90	45/86	37/82	62/88	75/95
Qwen3.5-Flash	59/92	69/93	65/83	74/94	80/99
best judge	71.5	59.8	51.2	56.4	70.2
mean of the ratings, no labels	75.3	63.9	56.7	59.2	71.2
median of the ratings, no labels	73.8	57.3	50.3	54.4	69.6
best reward model	61.7	50.3	44.0	63.8	75.0
pooling's regret at 8 prompts	2.4	4.8	2.4	2.3	2.2

Table 60: The addendum to the fifth protocol on WMT25's LLM-judge sets, configuration by configuration: best-of-n recovery on the evaluation half, in points, the mean over the fourth protocol's 300 draws, of pooling's pick with eight labelled segments (r_T at 8), of the label-free mean of the thirteen judges' standardised scores, and of the in-sample best judge, the judge with the highest recovery over all segments, named. Described after the grading: the judges that pick worse than chance over all segments, of thirteen, and the median judge's recovery. The last two columns are E1 and E2 by configuration; below each set, their means over its configurations with the protocol's interval, the percentile interval and the grade (`scripts/decision/ensemble_wmt25_r18.py`, frozen; `ensemble_wmt25_described_r18.py`).

pair	label	below chance	median judge	r_T at 8	mean	best judge	named	E1	E2
en-ko_KR	MQM	1	+29.6	+34.2	+55.3	+43.2	Claude-4	+21.1	+12.1
ja-zh_CN	MQM	1	+15.6	+28.8	+39.1	+48.5	Gemini2_5-Pro	+10.3	−9.3
mean	MQM			+31.5	+47.2	+45.8		+15.7	+1.4
							the protocol's interval	[+6.3, +25.1]	[−9.1, +11.9]
							percentile interval	[+6.7, +25.5]	[−11.7, +9.1]
							grade	supported	consistent
cs-de_DE	ESA	0	+16.1	+39.2	+45.0	+45.4	Gemini2_5-Pro	+5.8	−0.4
cs-uk_UA	ESA	2	+22.0	+35.2	+41.3	+46.9	Gemini2_5-Pro	+6.1	−5.6
en-ar_EG	ESA	9	−6.7	+31.5	−11.9	+38.5	GPT-4_1	−43.4	−50.3
en-bho_IN	ESA	8	−39.4	+44.6	−18.7	+54.4	GPT-4_1	−63.4	−73.1
en-cs_CZ	ESA	0	+16.9	+34.8	+39.0	+43.4	Gemini2_5-Pro	+4.2	−4.4
en-et_EE	ESA	2	+6.2	+41.0	+33.3	+45.4	Gemini2_5-Pro	−7.7	−12.0
en-is_IS	ESA	0	+10.7	+55.8	+52.9	+56.8	GPT-4_1	−2.9	−3.9
en-it_IT	ESA	1	+14.3	+25.2	+26.8	+33.9	Gemini2_5-Pro	+1.6	−7.1
en-ja_JP	ESA	3	+9.2	+18.0	+15.8	+25.0	Gemini2_5-Pro	−2.2	−9.2
en-ru_RU	ESA	1	+15.4	+24.8	+36.6	+34.1	Gemini2_5-Pro	+11.8	+2.5
en-sr_Cyrl_RS	ESA	1	+25.5	+38.7	+46.4	+54.4	Gemini2_5-Pro	+7.7	−8.0
en-uk_UA	ESA	0	+9.1	+18.0	+27.6	+25.9	GPT-4_1	+9.6	+1.7
en-zh_CN	ESA	1	+17.7	+20.1	+34.7	+33.5	DeepSeek-V3	+14.6	+1.2
mean	ESA			+32.9	+28.4	+41.4		−4.5	−13.0
							the protocol's interval	[−7.8, −1.2]	[−16.2, −9.7]
							percentile interval	[−6.7, +0.02]	[−16.3, −10.1]
							grade	contradicted	contradicted

Table 61: The second addendum to the fifth protocol: the jury of the three judges that pooled correlation ranks highest on eight labelled units, their scores averaged, against pooling's single pick with the same units (r_T at 8) and the label-free mean of all the judges (seven on PPE, thirteen on WMT25), in points: best-of-16 accuracy on the measurement set on PPE, best-of-n recovery on the evaluation half on WMT25, the mean over the protocols' draws (500 on PPE, 300 on WMT25). J1 and J2 are the jury minus pooling's pick and minus the mean of all; below each set, their means with the protocol's interval, the percentile interval and the grade. Right of them, described after the grading and not graded: juries of two, four and five ranked by pooled correlation, the jury of three ranked by the grouped mean, and pooling's pick and the jury of three at half the pool or half the segments (`scripts/decision/jury_r18.py`, frozen).

		graded, eight labelled units					described					
set		r_T at 8	jury of 3	mean of all	J1	J2	jury of 2	of 4	of 5	of 3 by grp	r_T at half	jury at half
MMLU-Pro	PPE	+69.1	+72.9	+75.3	+3.8	−2.4	+71.6	+73.8	+74.2	+72.9	+70.5	+73.1
MATH	PPE	+55.0	+60.7	+63.9	+5.7	−3.2	+58.7	+61.9	+63.1	+60.2	+56.4	+60.7
GPQA	PPE	+48.8	+52.6	+56.7	+3.7	−4.1	+51.2	+53.7	+54.6	+52.7	+49.0	+53.3
IFEval	PPE	+54.1	+57.2	+59.2	+3.1	−2.1	+56.0	+58.0	+58.4	+57.4	+54.3	+58.2
MBPP-Plus	PPE	+67.9	+70.8	+71.2	+2.8	−0.4	+69.9	+71.2	+71.2	+70.9	+68.7	+70.8
mean	PPE	+59.0	+62.8	+65.3	+3.8	−2.4	+61.5	+63.7	+64.3	+62.8	+59.8	+63.2
				the protocol's interval	[+3.0, +4.6]	[−3.6, −1.3]						
				percentile interval	[+3.0, +4.7]	[−3.6, −1.4]						
				grade	supported	contradicted						
en-ko_KR	WMT25 MQM	+34.2	+46.2	+55.3	+12.0	−9.1	+42.1	+48.3	+50.5	+50.4	+42.4	+50.1
ja-zh_CN	WMT25 MQM	+28.8	+44.9	+39.1	+16.1	+5.8	+41.1	+45.5	+44.4	+49.9	+43.6	+54.6
mean	WMT25 MQM	+31.5	+45.5	+47.2	+14.0	−1.7	+41.6	+46.9	+47.4	+50.1	+43.0	+52.4
				the protocol's interval	[+8.1, +20.0]	[−11.0, +7.7]						
				percentile interval	[+8.3, +20.1]	[−10.5, +6.7]						
				grade	supported	not supported						
cs-de_DE	WMT25 ESA	+39.2	+45.6	+45.0	+6.4	+0.6	+44.4	+46.6	+46.6	+44.8	+44.8	+46.5
cs-uk_UA	WMT25 ESA	+35.2	+44.0	+41.3	+8.8	+2.7	+42.3	+45.2	+45.5	+42.8	+46.0	+47.5
en-ar_EG	WMT25 ESA	+31.5	+30.7	−11.9	−0.8	+42.6	+34.5	+26.3	+22.4	+32.1	+38.3	+35.0
en-bho_IN	WMT25 ESA	+44.6	+67.5	−18.7	+22.9	+86.3	+63.2	+68.6	+67.7	+67.6	+44.4	+69.4
en-cs_CZ	WMT25 ESA	+34.8	+44.2	+39.0	+9.4	+5.2	+41.5	+45.1	+45.4	+44.4	+42.5	+47.9
en-et_EE	WMT25 ESA	+41.0	+44.0	+33.3	+3.0	+10.7	+44.1	+43.2	+42.2	+43.1	+45.4	+48.2
en-is_IS	WMT25 ESA	+55.8	+58.4	+52.9	+2.6	+5.5	+59.1	+58.7	+58.8	+58.5	+56.2	+58.4
en-it_IT	WMT25 ESA	+25.2	+31.1	+26.8	+5.9	+4.4	+29.9	+30.9	+30.7	+31.8	+31.6	+34.8
en-ja_JP	WMT25 ESA	+18.0	+22.1	+15.8	+4.1	+6.3	+21.4	+21.4	+20.4	+23.0	+20.4	+25.1
en-ru_RU	WMT25 ESA	+24.8	+34.8	+36.6	+9.9	−1.9	+31.5	+36.6	+37.5	+34.5	+29.4	+35.9
en-sr_Cyrl_RS	WMT25 ESA	+38.7	+51.2	+46.4	+12.5	+4.8	+47.5	+53.2	+54.7	+54.1	+46.8	+57.3
en-uk_UA	WMT25 ESA	+18.0	+24.2	+27.6	+6.2	−3.4	+22.7	+24.9	+25.6	+23.1	+21.4	+24.2
en-zh_CN	WMT25 ESA	+20.1	+29.5	+34.7	+9.4	−5.2	+26.5	+31.3	+32.3	+31.1	+24.1	+34.0
mean	WMT25 ESA	+32.9	+40.6	+28.4	+7.7	+12.2	+39.1	+40.9	+40.8	+40.8	+37.8	+43.4
				the protocol's interval	[+6.3, +9.2]	[+9.2, +15.2]						
				percentile interval	[+6.6, +9.5]	[+8.8, +14.5]						
				grade	supported	supported						

Table 62: The jury of the second addendum against a jury of three drawn at random, at the same three judge calls a response, read after the addenda and graded by none. r_T at 8: pooling's pick with eight labelled units; jury: the three judges pooling ranks highest on them, averaged; random three: the mean over every set of three judges of the panel of that three's average, which reads no label. Set means over the configurations (Cfg) in points, best-of-16 accuracy on PPE's measurement set, best-of-n recovery on WMT25's evaluation half, on the draws of the addendum; differences with the addendum's interval, its replicates reused. The indented row keeps the 860 of PPE's 960 measurement prompts whose question does not also stand in the pool. `scripts/decision/jury_baselines_r22.py`.

Set	Cfg	r_T at 8	jury	random three	jury − r_T	jury − random three	random three − r_T
PPE, seven LLM judges	5	59.0	62.8	62.6	+3.8 [+3.0, +4.6]	+0.2 [−0.3, +0.7]	+3.6 [+2.8, +4.5]
measurement prompts whose question is not in the pool	5	58.5	62.2	62.1	+3.7 [+2.9, +4.6]	+0.2 [−0.3, +0.6]	+3.6 [+2.7, +4.5]
WMT25 MQM, thirteen LLM judges	2	31.5	45.5	36.7	+14.0 [+8.1, +20.0]	+8.8 [+2.8, +14.9]	+5.2 [−0.5, +10.9]
WMT25 ESA, thirteen LLM judges	13	32.9	40.6	20.7	+7.7 [+6.3, +9.2]	+19.8 [+17.9, +21.8]	−12.1 [−14.3, −10.0]

Table 63: The benchmarks of the census, one row per normalised release or part. Labels and clusters count the numeric human labels and the cluster variables the item table names; configurations is their product, and estimable the configurations that enter the summaries of Figure 4 and Appendix W: at least five clusters, which 143 configurations have, and a typical cluster holding more than one item, which removes 5 more, the sentence-clustered configurations of speechocean762, where almost every sentence is read by one learner. Table 3 takes one configuration per benchmark and label on the item set its published numbers use (Appendix F); the decision experiment takes the benchmarks that release evaluator scores (Appendix H). FRANK is normalised once and counted three times, whole and by its two source corpora, and the three WMT language pairs are one benchmark in every count of the text. Versions and checksums of every source file are in the release's benchmark cards.

Benchmark	Field	Source	Labels	Clusters	Config.	Estimable	Table 3	Decision
speechocean762	pronunciation	(Zhang et al., 2021)	5	4	20	10	yes	yes
BVCC (VoiceMOS 2022 main track)	MOS prediction	(Cooper & Yamagishi, 2021)	1	3	3	2	yes	no
SOMOS	MOS prediction	(Maniati et al., 2022)	2	4	8	8	yes	no
VCC2018	MOS prediction	(Lorenzo-Trueba et al., 2018)	2	7	14	10	yes	no
TTSDS2 listening test	MOS prediction	(Minixhofer et al., 2026)	3	3	9	6	yes	no
SummEval	summarization	(Fabbri et al., 2021)	8	4	32	16	yes	yes
RealSumm	summarization	(Bhandari et al., 2020)	2	3	6	4	yes	yes
FRANK	summarization	(Pagnoni et al., 2021)	12	3	36	24	yes	yes
FRANK, CNN/DM part	summarization	(Pagnoni et al., 2021)	4	2	8	8	no	no
FRANK, XSum part	summarization	(Pagnoni et al., 2021)	4	2	8	4	no	no
HANNA	story generation	(Chhun et al., 2022)	6	2	12	12	yes	yes
HelpSteer2	chat responses	(Wang et al., 2024c)	5	2	10	5	no	no
TL;DR axis ratings	summarization	(Stiennon et al., 2020)	5	3	15	15	no	no
WMT22 MQM en-de	translation	(Freitag et al., 2022)	1	5	5	4	yes	yes
WMT22 MQM en-ru	translation	(Freitag et al., 2022)	2	5	10	6	yes	yes
WMT22 MQM zh-en	translation	(Freitag et al., 2022)	1	5	5	4	yes	yes
all					201	138		

Table 64: Every primary configuration: the payoff on the evaluation half of the evaluator selected at half the clusters by the pooled coefficient, the grouped mean and the benchmark's own statistic, the paired difference of the grouped mean against pooling in points, and the share of draws on which the two select different evaluators (dis., %). Same draws as Table 5.

			Best of n, recovery %					Pairwise choice, accuracy %				
Configuration	C	Ev.	r_T	grp	own	grp$-r_T$	dis.	r_T	grp	own	grp$-r_T$	dis.
FRANK factuality	499	11	43.4	49.8	47.6	+6.4	33	73.6	76.8	75.7	+3.2	33
HANNA coherence	96	19	33.9	34.5	34.7	+0.6	23	64.8	65.3	65.1	+0.4	23
HANNA complexity	96	19	37.4	42.2	41.5	+4.8	94	67.7	71.1	70.6	+3.4	94
HANNA empathy	96	19	31.2	28.7	28.6	−2.5	36	63.8	62.7	63.0	−1.1	36
HANNA engagement	96	19	32.0	29.8	29.5	−2.3	76	63.3	63.4	63.5	+0.2	76
HANNA relevance	96	19	26.2	47.2	45.7	+21.0	94	62.3	65.5	65.2	+3.2	94
HANNA surprise	96	19	12.6	12.9	12.7	+0.3	80	59.4	60.8	60.5	+1.4	80
RealSumm LitePyramid	100	16	38.3	40.7	40.7	+2.4	74	69.7	70.5	70.5	+0.8	74
SummEval coherence	100	16	64.0	56.7	56.7	−7.3	54	74.0	73.2	73.2	−0.8	54
SummEval consistency	100	16	92.3	97.1	97.1	+4.8	99	81.7	85.7	85.7	+4.0	99
SummEval fluency	100	16	78.0	71.6	71.6	−6.3	44	76.2	76.3	76.3	+0.0	44
SummEval relevance	100	16	62.7	62.7	62.7	+0.0	0	74.6	74.6	74.6	+0.0	0
WMT22 en-de	1315	18	46.9	46.9	47.8	−0.0	65	68.4	68.0	68.3	−0.3	65
WMT22 en-ru	1315	17	50.1	51.1	51.8	+1.0	47	70.9	71.2	71.4	+0.2	47
WMT22 zh-en	1875	18	41.2	41.4	40.7	+0.2	41	66.4	66.6	66.1	+0.2	41
speechocean762 accuracy	125	9	52.8	52.8	52.8	+0.0	0	70.4	70.4	70.4	+0.0	0
speechocean762 fluency	125	9	43.5	46.4	43.5	+2.9	73	73.2	75.4	73.2	+2.2	73
speechocean762 prosody	125	9	45.4	50.1	45.4	+4.7	55	71.2	71.9	71.2	+0.7	55
speechocean762 total	125	9	54.3	54.4	54.3	+0.0	1	70.9	70.9	70.9	+0.0	1

Table 65: Which rule picks the better evaluator, on every configuration that meets the entry criterion of Appendix H, with none selected out, and with all six rules scored on both decisions. Pairwise choice is the accuracy of choosing the better of two items of one cluster over all pairs whose labels differ, a tie in the evaluator counting one half; best of n is the achievable label gain recovered by picking the item the evaluator ranks first in each cluster, n being that cluster's size, summed over clusters so that clusters are weighted by the gain available in them, and clusters with no gain available are skipped; an evaluator tying several items at the top of a cluster is credited with their mean label. In grp, a cluster whose label varies and whose evaluator is constant scores zero rather than being dropped, so every evaluator is averaged over the same clusters; Appendix H gives what the common dropping convention does instead. Held out splits the clusters in half, selects on one half and pays on the other, and gives the percentage of evaluator pairs where the rule picks the evaluator that is not better on the held-out half, over 500 splits per configuration. r_T is the pooled correlation, r_W the within-cluster one, grp the mean of the per-cluster Spearman correlations with equal weight per cluster, pdp the pairwise difference Pearson of DiIanni & Deutsch (2025), and dec the decision itself measured on the selection half, which is the reference a practitioner could use. own is the benchmark's current headline statistic (Appendix G), which for machine translation is the tie-calibrated pairwise accuracy taken inside the source segment that the task has headlined since 2023; the flattened segment-level Kendall it headlined in 2022 is scored beside it in Appendix G. A system's outputs on different documents are not alternatives to one another, so the system-clustered configurations are a robustness check and not part of the primary analysis. The oracles we construct are excluded throughout.

					Held out, pairwise choice						Held out, best of n					
Benchmark	Label	Cluster	Ev.	Cl.	r_T%	r_W%	grp%	pdp%	own%	dec%	r_T%	r_W%	grp%	pdp%	own%	dec%
Clusters a decision chooses between																
FRANK	factuality	document	11	499	29.4	15.9	15.9	14.6	29.7	15.0	46.4	36.2	35.6	36.0	33.1	40.4
HANNA	coherence	prompt	19	96	15.3	10.6	12.0	10.6	13.7	10.7	16.8	19.6	16.4	19.6	14.8	16.8
HANNA	complexity	prompt	19	96	14.1	8.2	6.8	8.2	7.7	6.8	16.0	14.1	11.3	14.1	11.3	13.6
HANNA	empathy	prompt	19	96	13.0	9.2	10.4	9.2	11.8	9.4	14.6	16.9	16.1	16.9	16.7	17.5
HANNA	engagement	prompt	19	96	16.4	9.4	10.8	9.4	11.5	9.8	17.3	17.7	17.0	17.7	16.7	17.6
HANNA	relevance	prompt	19	96	22.2	12.5	12.3	12.5	12.7	12.1	26.6	18.9	16.8	18.9	17.0	19.7
HANNA	surprise	prompt	19	96	19.2	11.8	12.4	11.8	13.6	12.4	22.3	22.7	22.4	22.7	22.6	23.8
RealSumm	LitePyramid	document	16	100	38.2	18.8	14.2	18.8	14.2	14.3	44.0	31.7	28.5	31.7	28.5	39.7
speechocean762	accuracy	speaker	9	125	20.4	14.1	18.9	14.1	20.4	18.9	20.5	18.8	21.1	18.8	20.5	22.2
speechocean762	fluency	speaker	9	125	22.2	24.5	7.8	24.5	22.2	7.8	43.4	43.3	38.8	43.3	43.4	65.3
speechocean762	prosody	speaker	9	125	23.0	26.3	11.3	26.3	23.0	12.0	36.2	39.2	26.2	39.2	36.2	44.8
speechocean762	total	speaker	9	125	20.5	16.1	15.1	16.1	20.5	14.3	25.8	18.5	22.9	18.5	25.8	26.8
SummEval	coherence	document	16	100	11.1	9.9	10.5	9.9	10.5	9.5	16.8	15.3	17.6	15.3	17.6	14.8
SummEval	consistency	document	16	100	12.5	9.5	8.3	9.5	8.3	7.3	21.5	23.7	21.0	23.7	21.0	22.7
SummEval	fluency	document	16	100	14.0	13.5	12.0	13.5	12.0	10.8	24.8	26.9	24.9	26.9	24.9	20.4
SummEval	relevance	document	16	100	22.3	19.3	20.3	19.3	20.3	20.2	20.4	28.5	26.8	28.5	26.9	22.4
WMT22 MQM	ende	source segment	18	1315	17.0	13.6	14.2	13.6	20.4	8.5	37.6	33.6	23.8	33.6	41.2	15.3
WMT22 MQM	enru	source segment	17	1315	18.1	11.9	5.8	11.9	14.7	4.7	32.7	25.8	18.2	25.8	29.5	15.4
WMT22 MQM	zhen	source segment	18	1875	27.9	15.2	7.6	15.2	20.5	7.2	34.9	25.3	16.3	25.3	28.6	16.1
Clusters of one system, robustness only																
FRANK	factuality	system	11	9	49.9	18.1	14.1	18.1	18.1	12.3	76.2	38.1	37.7	38.1	38.1	33.4
HANNA	coherence	system	19	10	24.8	21.6	22.7	21.6	29.0	23.0	39.6	39.1	39.0	39.1	37.1	47.5
HANNA	complexity	system	19	10	19.8	18.9	19.4	18.9	24.7	19.8	41.4	39.7	39.6	39.7	39.1	45.0
HANNA	empathy	system	19	10	16.9	16.3	15.1	16.3	18.4	15.6	32.1	32.4	32.6	32.4	32.7	38.6
HANNA	engagement	system	19	10	20.9	20.2	20.5	20.2	26.1	20.7	44.5	44.1	43.5	44.1	42.6	50.3
HANNA	relevance	system	19	10	15.8	15.2	12.6	15.2	24.0	13.0	37.5	35.5	36.9	35.5	38.3	43.8
HANNA	surprise	system	19	10	29.7	34.6	31.9	34.6	33.0	31.7	39.2	38.0	35.4	38.0	41.2	39.9
RealSumm	LitePyramid	system	16	24	27.6	26.8	16.2	26.8	51.6	16.1	33.0	31.3	26.6	31.3	49.8	36.0
SummEval	coherence	system	16	16	20.6	20.0	19.0	20.0	25.4	19.9	27.5	29.3	30.1	29.3	29.1	37.7
SummEval	consistency	system	16	16	24.0	24.6	18.7	24.6	25.2	19.4	45.4	41.7	41.2	41.7	45.9	45.9
SummEval	fluency	system	16	16	20.2	19.7	15.7	19.7	24.2	17.0	37.6	38.7	35.8	38.7	35.2	33.4
SummEval	relevance	system	16	16	26.5	26.2	28.4	26.2	32.1	28.2	30.7	32.5	33.1	32.5	39.8	28.5
WMT22 MQM	ende	system	18	13	11.2	11.3	6.1	11.3	14.8	3.4	45.3	45.1	45.7	45.1	49.6	25.6
WMT22 MQM	enru	system	17	15	10.7	11.5	5.7	11.5	10.1	3.5	39.9	40.1	38.9	40.1	42.7	7.8
WMT22 MQM	zhen	system	18	14	8.0	7.8	4.4	7.8	12.4	4.0	47.8	48.1	48.5	48.1	48.8	45.3

Table 66: Which evaluator each rule selects, on every primary configuration and decision, over the 500 document-blocked splits: the evaluator named most often on the selection half by the pooled coefficient, the grouped mean and the benchmark's own statistic, with in parentheses the top-1 regret of that rule as a percentage of the span between the best evaluator on the held-out half and the average one, averaged over splits. Best is the evaluator most often best on the held-out half. Names are abbreviated; the full names and the family each belongs to are in `results/winners.json`.

Benchmark	Label	Decision	Ev.	Best on held-out	r_T names (regret%)	grp names (regret%)	own names (regret%)
FRANK	factuality	best-n	11	BertScore F1 Art	BertScore F1 Art (39)	BertScore F1 Art (0)	BertScore F1 Art (13)
FRANK	factuality	pair	11	BertScore F1 Art	BertScore F1 Art (28)	BertScore F1 Art (0)	BertScore F1 Art (8)
HANNA	coherence	best-n	19	Beluga-13B	OrcaPlatypus (20)	OrcaPlatypus (16)	OrcaPlatypus (17)
HANNA	coherence	pair	19	OrcaPlatypus	OrcaPlatypus (12)	OrcaPlatypus (5)	OrcaPlatypus (11)
HANNA	complexity	best-n	19	METEOR	Beluga-13B (21)	METEOR (11)	METEOR (12)
HANNA	complexity	pair	19	METEOR	Beluga-13B (29)	METEOR (3)	METEOR (7)
HANNA	empathy	best-n	19	OrcaPlatypus	Beluga-13B (24)	Beluga-13B (31)	Beluga-13B (31)
HANNA	empathy	pair	19	Beluga-13B	Beluga-13B (9)	Beluga-13B (18)	Beluga-13B (14)
HANNA	engagement	best-n	19	OrcaPlatypus	OrcaPlatypus (22)	METEOR (31)	METEOR (31)
HANNA	engagement	pair	19	BARTScore-SH	OrcaPlatypus (27)	METEOR (26)	METEOR (24)
HANNA	relevance	best-n	19	BARTScore-SP	OrcaPlatypus (63)	BARTScore-SP (7)	BARTScore-SP (10)
HANNA	relevance	pair	19	BARTScore-SP	OrcaPlatypus (35)	BARTScore-SP (5)	BARTScore-SP (8)
HANNA	surprise	best-n	19	OrcaPlatypus	Beluga-13B (56)	METEOR (54)	METEOR (54)
HANNA	surprise	pair	19	METEOR	Beluga-13B (34)	METEOR (16)	METEOR (19)
RealSumm	LitePyramid	best-n	16	bart_cnn_hypo_ref	bart_para_hypo_ref (56)	bart_cnn_hypo_ref (33)	bart_cnn_hypo_ref (33)
RealSumm	LitePyramid	pair	16	bart_cnn_hypo_ref	bart_para_hypo_ref (21)	bart_cnn_hypo_ref (9)	bart_cnn_hypo_ref (9)
speechocean762	accuracy	best-n	9	hmamba	hmamba (4)	hmamba (4)	hmamba (4)
speechocean762	accuracy	pair	9	hmamba	hmamba (0)	hmamba (0)	hmamba (0)
speechocean762	fluency	best-n	9	HierCB	hmamba (87)	WavLM+mean (66)	hmamba (87)
speechocean762	fluency	pair	9	WavLM+mean	hmamba (39)	WavLM+mean (2)	hmamba (39)
speechocean762	prosodic	best-n	9	WavLM+mean	hmamba (58)	WavLM+mean (22)	hmamba (58)
speechocean762	prosodic	pair	9	WavLM+mean	hmamba (27)	WavLM+mean (10)	hmamba (27)
speechocean762	total	best-n	9	hmamba	hmamba (9)	hmamba (9)	hmamba (9)
speechocean762	total	pair	9	hmamba	hmamba (0)	hmamba (0)	hmamba (0)
SummEval	coherence	best-n	16	G-Eval	G-Eval (2)	UniEval (29)	UniEval (29)
SummEval	coherence	pair	16	G-Eval	G-Eval (2)	UniEval (8)	UniEval (8)
SummEval	consistency	best-n	16	G-Eval	UniEval (15)	G-Eval (2)	G-Eval (2)
SummEval	consistency	pair	16	G-Eval	UniEval (22)	G-Eval (0)	G-Eval (0)
SummEval	fluency	best-n	16	UniEval	G-Eval (24)	G-Eval (42)	G-Eval (42)
SummEval	fluency	pair	16	UniEval	G-Eval (7)	G-Eval (7)	G-Eval (7)
SummEval	relevance	best-n	16	G-Eval	G-Eval (0)	G-Eval (0)	G-Eval (0)
SummEval	relevance	pair	16	G-Eval	G-Eval (0)	G-Eval (0)	G-Eval (0)
WMT22 MQM	ende	best-n	18	MetricX-xl	MetricX-xxl (19)	MetricX-xl (19)	MetricX-xl (14)
WMT22 MQM	ende	pair	18	MetricX-xxl	MetricX-xxl (5)	MetricX-xl (11)	MetricX-xl (6)
WMT22 MQM	enru	best-n	17	MetricX-xxl	MetricX-xxl (17)	MetricX-xxl (10)	MetricX-xxl (8)
WMT22 MQM	enru	pair	17	MetricX-xxl	MetricX-xxl (8)	MetricX-xxl (4)	MetricX-xxl (1)
WMT22 MQM	zhen	best-n	18	BLEURT-20	COMET-22 (21)	COMET-22 (20)	MetricX-xl (27)
WMT22 MQM	zhen	pair	18	COMET-22	COMET-22 (8)	COMET-22 (3)	MetricX-xl (18)

Table 67: The three WMT22 MQM source-segment configurations under three units of held-out split, the percentage of evaluator pairs a rule computed on the selection side orders against the decision on the measurement side. Segment: random halves of the source segments, whole segments at a time, the unit of the held-out experiment of Appendix H. Document: random halves of the source documents, whole documents at a time, the unit of Table 5. Rater: random halves of the MQM raters, whole raters at a time; on en-de and en-ru one rater annotates a whole document, so this moves whole documents as well, while on zh-en raters are assigned per document and system and 576 segments carry several raters and are placed with their most frequent one, so that split is rater-disjoint only approximately (Appendix I); 500 splits under each of the three, from one seed. Domain: the rule is computed on three of the test set's four domains and scored on the fourth, over the four, with no randomness. Segments / documents gives the cluster and document counts. own is the tie-calibrated pairwise accuracy inside the source segment.

			Pairwise choice					Best of n				
Pair	Seg. / doc.	Split unit	r_T%	r_W%	grp%	own%	dec%	r_T%	r_W%	grp%	own%	dec%
en-de	1315 / 181	segment	16.5	13.4	13.6	20.1	7.9	37.6	34.0	23.5	41.4	15.2
		document	17.0	13.6	14.2	20.4	8.5	37.6	33.6	23.8	41.2	15.3
		rater	17.0	13.7	13.1	20.3	7.6	36.6	32.6	22.2	40.3	14.2
		domain, leave one out	19.5	17.7	21.4	21.5	17.0	40.4	38.6	31.9	42.4	22.2
en-ru	1315 / 181	segment	17.6	11.7	5.5	14.5	4.4	32.5	25.7	18.0	29.4	14.6
		document	18.1	11.9	5.8	14.7	4.7	32.7	25.8	18.2	29.5	15.4
		rater	18.8	12.8	10.3	15.7	8.6	31.0	23.2	18.8	28.5	19.8
		domain, leave one out	20.7	13.8	10.9	16.1	10.5	29.4	21.9	18.2	25.4	20.0
zh-en	1875 / 256	segment	27.6	15.0	7.2	20.6	6.7	34.5	25.4	16.1	28.7	15.0
		document	27.9	15.2	7.6	20.5	7.2	34.9	25.3	16.3	28.6	16.1
		rater	27.9	15.5	8.7	20.0	7.8	36.3	27.2	18.0	29.8	17.8
		domain, leave one out	28.6	17.5	11.8	21.9	11.3	39.1	31.9	26.8	35.0	25.5
all three		segment	20.5	13.4	8.8	18.4	6.3	34.9	28.4	19.2	33.2	15.0
		document	21.0	13.6	9.2	18.5	6.8	35.1	28.2	19.4	33.1	15.6
		rater	21.2	14.0	10.7	18.7	8.0	34.7	27.6	19.6	32.9	17.3
		domain, leave one out	22.9	16.3	14.7	19.8	12.9	36.3	30.8	25.6	34.2	22.6

Table 68: speechocean762 test set, 125 speakers with 20 utterances each. η_y^2 is the between-speaker share of label variation. The population ceiling is $\sqrt{\hat{\rho}}$ with a 95% interval from 2000 bootstraps of the speakers. Profile columns are pooled correlations of predictors that use only other utterances of the speaker, averaged over 20 split-half partitions. WavLM ridge is fit on the scored utterance and evaluated on all test utterances. The last column is the within-speaker correlation a perfect item-level predictor would reach against the released labels, the product of the *square root* of the within-speaker reliability of the rater mean and the correlation of the released label with that mean, the released label being a median or a floored median rather than that mean; simulation on the real panel puts it within about 0.03 of the exact value, so it is a guide rather than a strict bound.

				Item-blind		WavLM ridge		
Aspect	η_y^2	Population ceiling	Best published	Labels	WavLM	r_T	r_W	Rater-limited r_W
Accuracy	0.66	0.805 [0.762, 0.838]	0.807	0.785	0.713	0.726	0.246	0.724
Fluency	0.73	0.846 [0.806, 0.880]	0.848	0.830	0.763	0.799	0.421	0.720
Prosody	0.75	0.861 [0.827, 0.888]	0.843	0.847	0.776	0.791	0.355	0.683
Total	0.70	0.827 [0.786, 0.858]	0.830	0.809	0.737	0.739	0.271	0.744

Table 69: Pooled correlation r_T and within-speaker correlation r_W on the speechocean762 test set for evaluators that use the mean WavLM feature of the speaker's test utterances. History pooling is Equation 4 applied to the plain ridge output, with its three coefficients cross-fitted over held-out test speakers. Concatenation and the two-component form receive the speaker mean feature as input. MLP rows average five seeds, whose standard deviations are at most 0.003.

		Accuracy		Fluency		Prosody		Total	
		r_T	r_W	r_T	r_W	r_T	r_W	r_T	r_W
Ridge	Plain	0.726	0.246	0.799	0.421	0.791	0.355	0.739	0.271
	History pooling	0.743	0.244	0.812	0.416	0.808	0.348	0.757	0.268
	Concatenation	0.749	0.287	0.818	0.446	0.816	0.382	0.770	0.289
	Two-component	0.745	0.279	0.816	0.443	0.815	0.381	0.768	0.275
MLP	Plain	0.724	0.240	0.800	0.410	0.793	0.347	0.741	0.247
	Concatenation	0.748	0.270	0.818	0.435	0.818	0.370	0.770	0.275
	Two-component	0.739	0.275	0.807	0.431	0.806	0.372	0.759	0.272

Table 70: Reproduced evaluators on the speechocean762 test set, total score. Published is the value reported by the system's own paper or release, and the WavLM ridge regression is our own baseline, so it has none. Every row is one trained model. Retrained rows use the last epoch of the first seed, with the standard deviation of the pooled correlation across three seeds in parentheses and, in brackets, the 2.5 and 97.5 percentiles of 2000 bootstraps of the speakers; averaging the three seeds' predictions instead raises the pooled correlation of HMamba to 0.834. The text column is the within-speaker correlation after the text-only prediction is partialled out, and the last column is the pooled correlation after history pooling. *selected on the test set by its authors. †one released checkpoint against a five-seed published mean. ‡published value from the first arXiv version (2406.02859v1) of Yan et al. (2025c), a preprint that its second arXiv version withdraws; the value's epoch-selection rule is not documented. The retrained rows have no development set, as in the released training code, so they report the last epoch rather than a selected one.

Evaluator	Published r_T	r_T	r_W	r_B	r_W without text	r_T after pooling
WavLM ridge	–	0.739	0.271 [0.195, 0.345]	0.893	–	–
GOPT, released checkpoint*	0.743	0.744	0.379 [0.299, 0.454]	0.922	0.347	0.794
GOPT, retrained	0.742	0.740 (0.008)	0.383 [0.305, 0.458]	0.919	0.351	0.793
MultiPA, released checkpoint†	0.761	0.746	0.304 [0.240, 0.373]	0.905	0.276	0.769
HierCB with ConPCO, retrained‡	0.803	0.794 (0.004)	0.422 [0.325, 0.514]	0.940	0.393	0.815
HMamba, retrained	0.829	0.824 (0.002)	0.479 [0.413, 0.542]	0.959	0.456	0.841

Table 71: Paired differences on the speechocean762 total score with 125 test speakers. MDE is the minimum detectable difference at 80% power and 5% two-sided level from the paired cluster bootstrap of 1000 resamples. Speakers is the number of test speakers needed for an MDE of 0.01, extrapolated with the $1/\sqrt{C}$ scaling that resampling 25 to 500 speakers matches to within 8%. The MDE column uses the paired-bootstrap standard error at 125 speakers and the Speakers column the standard error of the same comparison in the resampling curve at its 125-speaker point, which differ by about 3%.

	Pooled r_T			Within-speaker r_W		
Comparison	Difference	MDE	Speakers	Difference	MDE	Speakers
GOPT against WavLM ridge	0.005	0.103	13,900	0.107	0.201	47,900
WavLM concatenation against plain	0.031	0.023	700	0.017	0.024	800
Two GOPT training seeds	0.002	0.013	200	0.014	0.026	800

Table 72: Label-only statistics for every benchmark and label analysed, each row on the item set its benchmark's published numbers use and in the coefficient they use. C is the number of clusters, starred below twenty. Share is the between-cluster share of the label variance the protocol's correlation sees. Bound is what no cluster-constant predictor exceeds under that protocol, a quantity of the released labels and not a claim that cluster identity by itself supplies it: the exact in-sample maximum under a Pearson or a per-row protocol (Propositions 1 and 4), and under a pooled rank protocol, marked §, a certified upper bound (Proposition 5), which Table 2 reports beside the best value a search attains. Oracle is the leave-one-out label oracle of Proposition 3, which reads the human labels of the cluster's other items and never the item it scores, with the 2.5 and 97.5 percentiles of a cluster bootstrap of 400 resamples (300 where raters are crossed with items), an interval Section 3 reports as descriptive because it under-covers at every cluster count simulated. Best published is the strongest published result for that label, read from its primary source; Ratio is the oracle over it, printed only where the item set, aggregation and coefficient were confirmed to match. ‡ marks a number matching on aggregation and coefficient but not, or not confirmably, on item set, its ratio in parentheses; † one on a different protocol, which gets none. Rows are grouped by domain. Appendix F documents every row and Appendix G the level at which each benchmark's own papers validate evaluators.

Benchmark	Label	Cluster	Protocol	C	Share	Bound	Oracle [boot. 2.5–97.5%]	Best / ratio
Speech assessment and MOS prediction								
speechocean762	accuracy	speaker	utt. Pearson	125	0.66	0.815	0.793 [.74, .83]	0.807 / 0.98
speechocean762	fluency	speaker	utt. Pearson	125	0.73	0.854	0.836 [.79, .87]	0.848 / 0.99
speechocean762	prosodic	speaker	utt. Pearson	125	0.75	0.868	0.852 [.81, .88]	0.843 / 1.01
speechocean762	total	speaker	utt. Pearson	125	0.70	0.835	0.816 [.77, .85]	0.830 / 0.98
BVCC	mos	system	utt. Spearman	187	0.77	0.876§	0.782 [.70, .82]	0.897 / 0.87
SOMOS	mos	system	utt. Spearman	201	0.34	0.577§	0.479 [.39, .52]	0.681‡ / (0.70)
VCC2018	mos	system	utt. Spearman	26	0.36	0.564§	0.476 [.25, .58]	0.718 / 0.66
TTSDS2	mos	system	utt. Pearson†	20	0.33	0.571	0.516 [.32, .62]	–
Summarization and story generation								
SummEval	coherence	system	doc. Spearman	16*	0.35	0.567	0.555 [.52, .59]	0.652 / 0.85
SummEval	consistency	system	doc. Spearman	16*	0.36	0.566	0.474 [.44, .51]	0.655 / 0.72
SummEval	fluency	system	doc. Spearman	16*	0.27	0.443	0.379 [.33, .43]	0.571 / 0.66
SummEval	relevance	system	doc. Spearman	16*	0.23	0.458	0.447 [.39, .50]	0.593 / 0.75
RealSumm	LitePyramid	system	doc. Spearman	24	0.11	0.309	0.298 [.23, .36]	0.53‡ / (0.56)
HANNA	coherence	system	prompt Kendall	10*	0.19	0.339	0.291 [.23, .36]	0.253 / 1.15
HANNA	relevance	system	prompt Kendall	10*	0.06	0.187	0.143 [.09, .23]	0.300 / 0.48
HANNA	empathy	system	prompt Kendall	10*	0.14	0.278	0.230 [.19, .30]	0.236 / 0.97
HANNA	surprise	system	prompt Kendall	10*	0.13	0.298	0.225 [.19, .32]	0.245 / 0.92
HANNA	engagement	system	prompt Kendall	10*	0.24	0.360	0.336 [.27, .39]	0.308 / 1.09
HANNA	complexity	system	prompt Kendall	10*	0.34	0.445	0.360 [.34, .45]	0.433 / 0.83
FRANK	factuality	system	pooled Pearson†	9*	0.67	0.820	0.819 [.63, .88]	0.47†
Machine translation and preference data								
WMT22 en-de	mqm	system	seg. Kendall	13*	0.01	0.105§	0.018 [.01, .03]	0.369 / 0.05
WMT22 en-de	mqm	segment	seg. Kendall	1315	0.59	0.709§	0.531 [.51, .55]	0.369 / 1.44
WMT22 zh-en	mqm	system	seg. Kendall	14*	0.04	0.151§	0.031 [.02, .04]	0.428 / 0.07
WMT22 zh-en	mqm	segment	seg. Kendall	1875	0.56	0.680§	0.515 [.50, .53]	0.428 / 1.20
WMT22 en-ru	mqm	system	seg. Kendall	15*	0.03	0.189§	0.092 [.08, .10]	0.420 / 0.22
WMT22 en-ru	mqm	segment	seg. Kendall	1315	0.63	0.702§	0.543 [.53, .56]	0.420 / 1.29
HelpSteer2	helpfulness	prompt	pooled Pearson	519	0.50	0.704	−0.009 [-.11, .09]	–
TL;DR axis	overall	policy	pooled Pearson	13*	0.21	0.460	0.457 [.29, .56]	–

Table 73: Coverage, in %, of the main run's interval for the paired best-of-n difference, the grouped mean's evaluator against pooling's, under the whole pipeline: 300 half-splits, the argmax that names an evaluator with the draw's tie-break, the ratio scored on the other half, and 200 corpus bootstraps of 300 splits, the estimate ± 1.96 standard deviations, over 200 replications per world and configuration (a coverage of 95% then has a standard error of 1.5 points), at eight clusters and at half. Worlds: real, the configuration's units as the population and a corpus drawn from them with replacement, its truth the mean over such corpora of the corpus's own estimand (300 corpora of 60 splits; the two truth columns, in points); null, the labels and clusters real, the panel replaced by exchangeable evaluators with fresh noise in every corpus, so the truth is zero; small effect, the null world with one evaluator's noise scaled by 0.9 (in the text); uneven sizes, the null world with a random half of the clusters cut to two items; cross, a panel built from the label's between- and within-cluster parts so that one evaluator ranks the clusters well and the candidates poorly and another the reverse, truth by Monte Carlo; heterogeneous null, two evaluators with the same within-cluster part and different between-cluster parts, so that pooling prefers one while every payoff difference between the two is zero in truth, the rest of the panel the cross world's fillers, run on the primary configurations without WMT and on the two validation sets with the panel capped at twenty, its truth by Monte Carlo, 200 corpora of 60 splits, since a rule may name one of the fillers. The other synthetic worlds run on the primary configurations without WMT. The summary rows add the percentile interval's coverage, the lowest coverage over configurations, the share of replications whose interval excludes zero, the range over the configurations on which the two rules disagree on at least a twentieth of the draws of the mean bootstrap standard deviation over the standard deviation of the estimate across replications, the mean coverage over the primary configurations alone, and the coverage of the interval of the mean over the primary configurations, its replicates resampled jointly within a benchmark.

	truth, real		real		null		hetero. null		cross		uneven sizes	
Configuration	8	half	8	half	8	half	8	half	8	half	8	half
so762 accuracy	−0.1	+0.0	98	100	100	98	96	100	94	95	100	100
so762 fluency	−0.1	+0.3	99	90	100	100	96	99	94	96	100	99
so762 prosodic	−2.1	+3.7	94	92	100	99	96	98	96	96	100	98
so762	+2.2	+0.1	92	100	100	99	98	98	95	94	98	98
FRANK factuality	+2.2	+7.9	92	87	99	98	96	96	98	88	100	100
HANNA coherence	+1.2	+1.1	96	98	100	98	100	99	100	98	100	98
HANNA complexity	+2.4	+3.4	98	90	100	100	98	100	100	98	100	98
HANNA empathy	−1.8	−2.1	100	99	100	100	98	100	100	99	100	100
HANNA engagement	+0.5	−1.0	98	99	100	100	100	99	100	100	99	99
HANNA relevance	+10.5	+18.5	90	81	99	99	99	99	100	100	100	100
HANNA surprise	+0.7	−0.1	98	94	100	99	100	100	100	99	100	100
RealSumm	+2.4	+2.7	98	96	99	98	95	96	98	100	100	100
SummEval coherence	−2.4	−6.0	92	80	100	98	98	100	97	100	100	100
SummEval consistency	+1.6	+4.4	92	90	100	99	99	100	100	100	100	100
SummEval fluency	−3.5	−5.8	96	88	99	99	100	99	100	100	99	96
SummEval relevance	−0.2	+0.0	98	100	100	100	97	96	94	100	100	99
WMT22 ende	+1.5	−0.2	92	98	–	–	–	–	–	–	–	–
WMT22 enru	+4.2	+0.9	96	96	–	–	–	–	–	–	–	–
WMT22 zhen	+0.3	−0.0	97	100	–	–	–	–	–	–	–	–
WMT23 ende	−0.5	−2.4	94	94	–	–	94	98	–	–	–	–
WMT23 heen	+1.5	−1.6	94	95	–	–	94	98	–	–	–	–
WMT23 zhen	+0.2	+4.7	94	92	–	–	96	98	–	–	–	–
RB2 factuality	−0.8	+1.2	96	98	–	–	96	100	–	–	–	–
RB2 precise if	+3.7	+0.0	92	100	–	–	100	100	–	–	–	–
RB2 math	−1.4	+5.7	98	98	–	–	98	100	–	–	–	–
RB2 safety	−4.8	+1.9	94	90	–	–	98	100	–	–	–	–
RB2 focus	−8.5	−0.6	92	100	–	–	96	100	–	–	–	–
mean coverage			95	94	100	99	97	99	98	98	100	99
mean coverage, percentile			97	97	100	100	98	97	98	98	100	100
lowest coverage			90	80	99	98	94	96	94	88	98	96
excluding zero			24.8	15.1	0.4	1.1	19.2	1.4	29.8	29.3	0.4	1.1
SD ratio, range (dis. $\geq$ 5%)			0.86–1.31	0.73–1.53	1.32–1.68	1.13–1.45	0.96–1.60	0.90–1.51	0.97–1.56	0.83–2.14	1.32–1.69	1.10–1.34
mean coverage, primary configurations			96	94	100	99	98	99	98	98	100	99
primary mean	+1.0	+1.5	98	93	–	–	–	–	–	–	–	–

Table 74: The other contrasts of Table 5 under the same pipeline in the real world alone, beside the calibrated one: the truth of the primary mean in points, the mean and lowest (low) coverage over the 27 configurations, the coverage of the interval of the mean over the primary (prim), RewardBench 2 and WMT23 configurations, and the primary mean's bootstrap standard deviation over the true one (ratio), at eight clusters and at half; 200 replications each; the bootstrap behind the rows, in their order, averages 300, 300, 300, 300 splits per resampled corpus.

Contrast	eight clusters							half the clusters						
	truth	mean	low	prim	RB2	WMT23	ratio	truth	mean	low	prim	RB2	WMT23	ratio
grouped mean vs pooling, best of n	+1.0	95	90	98	95	94	1.09	+1.5	94	80	93	95	95	0.95
pairwise accuracy vs pooling, best of n	+1.6	95	89	97	97	96	1.06	+2.0	95	81	92	96	96	0.95
grouped mean vs pooling, pairwise choice	+0.5	95	90	96	92	96	1.05	+1.0	94	82	92	96	89	0.99
best of n read in sample vs grouped mean, best of n	−4.0	96	90	98	92	94	1.26	+0.5	98	90	98	99	94	1.17