RAIM: ROBUST AGGREGATION OF INEXPENSIVE MODELS FOR HALLUCINATION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automatic evaluation of faithfulness increasingly relies on a large language model acting as a judge, yet the most reliable judges are proprietary frontier models. Their per-call cost, rate limits, and off-premises inference, however, make them ill-suited to the continuous, high-throughput monitoring that benchmarking and production use demand. In this work we investigate three fundamental questions: whether a panel of cheap open-weight judges (4–9B parameters) can be aggregated to stand in for a frontier one, what the substitution sacrifices in agreement (Cohen's κ) and accuracy, and when it is worth making. To address them, we propose RAIM, an aggregation scheme robust to the members' correlated errors, which couples: (a) a cross-fitted stacked logistic regression that fits the members' weights jointly, therefore penalising weaker judges echoing a stronger one; and (b) an admissibility test that, read from the members' own outputs, identifies a regime in which aggregating them improves on their best member and stays within reach of the frontier judge. We instantiate RAIM under ten judges drawn from disjoint families and we measure the substitution across eight faithfulness benchmarks. Read as paired differences against CLAUDE SONNET, the substitution clearly improves on one benchmark and clearly worsens on three (only two by a non-negligible margin), leaving the remaining four unresolved. The panel retains a median 93% of the frontier judge's κ agreement and gives up only 2.9 points of balanced accuracy on average. Inference runs at a sixty-fourth of the frontier's price at conventional cloud rates, or at the electricity cost on premises, so the operative expense is a one-time in-domain calibration of some fifty to a hundred labelled records, amortised over a few tens of thousands of evaluated items. Against the literature, the panel is competitive with purpose-trained detectors on their home benchmarks, losing 1.3 points of accuracy to `GPT-4o` and 1.9 to the best LLM-AGGREFACT leaderboard model; yet the strongest detector we reran falls 6 points behind on our grounded sets, and cannot score the ungrounded ones. In conclusion, whether pursuing the aggregation is worthwhile depends on the members themselves, making the gain auditable from how widely competence is spread across them and how far their errors decorrelate — both read at no further cost off the aggregator's calibration set. If several capable members err on different items, the panel improves on its best judge and approaches the frontier; conversely, where one dominates, the stacker recovers the leader, yet only there does the frontier remain materially ahead. A panel of cheap judges can therefore stand in for a frontier one at a small fraction of the cost wherever this audit admits it.

1 INTRODUCTION

Automatic evaluation of faithfulness –whether a given claim, a retrieved passage, or a model's output is supported by its source or constitutes a hallucination– increasingly relies on a large language model acting as a judge (LLM-as-a-judge; Zheng et al., 2023; Li et al., 2024a). Single-model judges, however, carry systematic biases (Panickssery et al., 2024; Chen & Goldfarb-Tarrant, 2025) and are reported to reach human-level agreement only at scales well above the small models ($<$10B parameters) a deployer can serve locally (Han et al., 2025). As a consequence, the judges one can most rely on are large proprietary models served from the cloud, so each evaluated item is a metered,

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.

rate-limited call, whose cost is small individually yet adds up at scale, and which sends the text off-premises to an endpoint the deployer does not control.

Three main strategies have been proposed to reduce this dependence: *escalate*, keeping a frontier judge but calling it only where a cheap model is unsure (Chen et al., 2024; Jung et al., 2025); *specialise*, training one small model expressly for the task (Tang et al., 2024; Kim et al., 2024b); or *aggregate*, running several cheap judges and combining their verdicts (Verga et al., 2024). Escalation, however, retains a frontier judge as a fallback, and we read it as a complementary solution (Appendix L). Specialists are typically trained for reference-conditioned verification and, whilst holding an advantage on home ground, they cannot score a task that supplies no reference at all. Moreover, they are reported to generalise poorly off the distribution they were trained on (Huang et al., 2025), as we confirm by running them ourselves across our suite (Appendix G). Specialisation and aggregation both remove the frontier model from the loop, as the constraints beyond price actually require (i.e. on-premises inference, no rate limit, no deprecation). Of the two, we pursue aggregation, the only one that needs no training of model weights and keeps the generality just noted. The practical question is therefore not whether a frontier judge is good, but whether one has to be paid for: can several cheap, open-weight judges, run locally and combined, stand in for it? We refer to Appendix L for a deeper account of these routes, of the noisy-voter literature the aggregator inherits, and of how recent panel studies compare with ours (Appendix A maps every appendix to the claim it supports).

The intuition that the *aggregate* solution might work is the classical one behind ensembling and the Condorcet jury theorem (de Condorcet, 1785), and a panel of small judges has indeed been reported to track human ratings better than a single frontier model at a fraction of its price (Verga et al., 2024). That intuition rests, however, on an assumption which is fragile for exactly these estimators. Weak voters rival a strong one when their errors are *independent*, whereas judges drawn from a common pretraining prior err together (Zhao et al., 2026; Hossain et al., 2026). Indeed, a panel of nine frontier judges is reported to carry only about two effective independent votes, leaving the best single judge unbeaten (Kohli, 2026). What this line of work leaves open is hence not if a cheap panel pays on average, but under which conditions it pays at all, and what the substitution costs where it does not.

We take up both questions together, and we do so by introducing RAIM (Robust Aggregation of Inexpensive Models; Figure 1), an aggregation scheme that couples two components. It aggregates a panel of cheap open-weight judges by (a) a cross-fitted stacked meta-learner, which fits member weights *jointly* and so discounts a judge echoing a stronger one, instead of weighting each by its own accuracy; and it pairs that with (b) an admissibility test, read off the single members, which identifies the regime in which aggregating them pays. Together, the two make the aggregation *robust*, where a plain ensemble is only redundant. We instantiate the scheme with ten small judges from disjoint families, and study its behaviour across eight faithfulness benchmarks and against a ladder of comparators ordered by the labelled information each of them consumes. Combining noisy voters is itself classical (Dawid & Skene, 1979; Wolpert, 1992), and we adopt the simplest form that expresses the joint weighting correlated errors require. A heavier estimator would consume labels the calibration budget cannot afford, and would offer a rival explanation for whatever the panel gains; with ten fitted coefficients, instead, the gain becomes attributable to the members alone.

The result proves conditional. A frontier judge is not, in general, replaced without cost, yet that cost is small and concentrated; against its own best member, conversely, the panel improves on all four core grounded sets, though the sample resolves that improvement on only one of them, with no clear worsening anywhere. The eight benchmarks therefore yield an account of *which* tasks fall on which side, with the members' joint distribution of *competence* and the *correlation* of their errors at the centre: aggregation pays where several judges are individually capable and disagree in their mistakes, and falls back on the leader where one judge dominates or all of them are weak and err together.

Contributions. We summarise our contributions as follows.

(i) We introduce RAIM, an aggregation scheme coupling (a) a cross-fitted meta-learner that fits the members' weights jointly with (b) a label-cheap admissibility test which, read off the members' own verdicts rather than the panel's, identifies a regime in which aggregating them pays. Along with it, we adopt a comparator ladder of baselines and a decision rule built on paired differences, which keeps the direction of an effect apart from the power to resolve it.

(ii) We instantiate RAIM on ten judges of 4–9B parameters from disjoint families over eight faithfulness benchmarks, releasing the whole chain from generation to analysis.

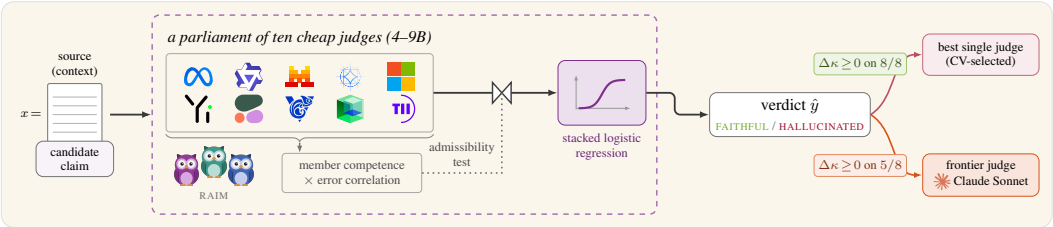


Figure 1: The RAIM protocol at a glance. Ten cheap open-weight judges (4–9B) each score a candidate against its source, and a cross-fitted stacked logistic regression combines their votes into one verdict. The same calibration set yields an admissibility test, read off the members, which identifies a regime in which aggregating them pays. The verdict is read against the CV-best single judge *of* the panel, and against a paid frontier judge the panel is meant to replace.

(iii) We measure what the panel trades off against a frontier judge, finding that it retains a median 93% of CLAUDE SONNET's κ and falls materially behind on two of the eight benchmarks. Against the purpose-trained detectors that are its sharpest competitors, it stays competitive on the benchmarks those detectors were trained for, and leads them elsewhere.

(iv) We price the substitution, including the labels the aggregator consumes, showing that inference runs at a sixty-fourth of the frontier's cost at rented cloud rates, so the operative expense is the one-time in-domain calibration and not the evaluation: fifty labelled records suffice for the aggregator to beat a free majority vote on seven of eight sets, and a hundred are enough to approach full supervision, which places the break-even against the frontier judge between 40,000 and 81,000 evaluated items.

(v) We identify what decides the outcome, and it is not the grounded-versus-ungrounded contrast one would expect but a property of the members: how widely competence is spread across them, and how far their errors decorrelate. We confirm this finding at the item level, over the 1,998 items on which the panel's own leader errs, and we show both coordinates to be readable from the calibration set the aggregator already consumes, so the regime is known by the time the panel is fitted.

2 THE RAIM PROTOCOL

Let us consider a panel made of N_v members (LLMs) which, given a prompt, return one binary verdict on a candidate statement: *supported* (gold$=0$) or *hallucinated*/unsupported (gold$=1$). We evaluate the panel on two tasks: *grounded* faithfulness, where the candidate is judged against an explicit reference; and *ungrounded* factuality, where it is judged against world knowledge alone. Datasets are built class-balanced, and items are grouped into clusters by the source record or grounding document they derive from (Appendix J). In what follows, we describe the RAIM protocol in two steps: the framework first and the instantiation after. We defer the description of the admissibility test to §3.3, where the quantities it reads are defined. The judge prompt, the decoding settings, and the coverage audit are deferred to Appendix K, and the per-dataset item construction to Appendix J.

Aggregation and the comparator ladder. Given N_v votes per query, we adopt a *stacked logistic regression* (Wolpert, 1992) to aggregate them: a vote over the votes, taking the member verdicts as its features and the gold label as its target, so that each member earns a weight from how well its verdict predicts the truth w.r.t. the others. Writing $v_i \in \{0, 0.5, 1\}$ for member i's verdict on an item ($i \in \{1, \ldots, N_v\}$; abstentions imputed as 0.5, tested in Appendix K), the meta-learner estimates

$$\mathbb{P}(y = 1 \mid v_1, \ldots, v_{N_v}) = \sigma\left(\beta_0 + \sum_{i=1}^{N_v} \beta_i v_i\right), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}, \tag{1}$$

fitting β by penalised maximum likelihood under an untuned ℓ_2 penalty[1]. The coefficients β_i are what a vote weighted by each member's own accuracy cannot express: they are fitted *jointly*, so a member that merely echoes a stronger one is discounted rather than counted twice, and a member that

[1]Of weight $1/C$ on the squared coefficients, left at the conventional $C = 1$; with only ten features the fit is insensitive to the choice, and the smallest sets (CNN, WICE, FACTSCORE) are where it matters most.

is informative but anti-correlated with gold can take a negative weight. To avoid leakage, the stacker is fitted by 5-fold cross-fitting, so that every item's stacked prediction is produced out-of-fold. A verdict is read at the conventional 0.5 threshold, which also breaks the unweighted majority's ties at $N_v = 10$. Fitting β, however, requires gold labels, a cost we measure in Appendix C and price in §4.

We compare the stacked panel against a ladder of comparators, ordered by the labelled information each of them consumes. The unweighted majority panel, the average member, and the unsupervised Dawid–Skene aggregator need no labels and are fully deployable; the best-on-average single is one global label-based selection deployed everywhere; the cross-validated (CV) best single, selected on a development split, and the stacked panel are both supervised and both evaluated strictly out-of-fold; the oracle best single, selecting the argmax on the very test set it is scored against, is the most generous yet undeployable comparator; and above them all sits the one the panel is meant to replace rather than beat, a proprietary frontier judge, named and read in §4.

Metrics and the decision rule. Our primary metric is Cohen's κ (Cohen, 1960), which rescales the observed agreement p_o with the gold label by the agreement p_e that a random sample would produce, so that only agreement above chance is credited: $\kappa = (p_o - p_e)/(1 - p_e)$. We accompany it with balanced accuracy (BACC), i.e. the mean of the two per-class recalls (on our class-balanced sets coinciding with plain accuracy), which accounts for an abstention from the opposite side: an abstained item is dropped from κ, yet charged in BACC as an error against its own class; Appendix K audits the denominator asymmetry this creates between an abstaining baseline and the always-imputing panel. All quantities carry 95% paired cluster-bootstrap percentile intervals over $B = 2000$ resamples of the clustering unit, so that co-clustered instances always resample together, and every central estimate is followed by its $[\text{lower}, \text{upper}]$ interval. When comparing two methods, every verdict is read from the paired difference of their scores: within each draw, both sides are recomputed on the same resampled clusters, so that scores are always compared on the very same items; this is materially sharper than assessing if the marginal intervals overlap, since the two scores rise and fall together across draws[2].

On that basis we define three verdicts: a *win* requires the stacked κ to significantly beat the CV-best single, i.e. the lower bound of the paired difference to exceed zero; a *clean win* additionally requires it to beat the oracle single; a *tie* is a difference straddling zero, and admits no intermediate grade, though where a tie's interval comes close to excluding zero we report its one-sided p. Every paired difference we report –that one included, and equally those against the average member, the label-free aggregators, or the frontier judge of §4– is described in one four-way form: an *improvement* or a *worsening*, *clear* where the 95% paired interval, uncorrected for multiplicity, excludes zero and *unresolved* where it does not (Appendix K gives Holm-adjusted p-values for the best-single contrasts).

Benchmarks and judges. We run RAIM on eight faithfulness benchmarks[3]: six *grounded* and two *ungrounded* (Appendix I, Table 14). Two of the six grounded sets are held back from the primary tally on measurability grounds: CNN is too small to power the paired contrast ($n = 114$), and EXPERTQA is a low-headroom task, with no published method on the LLM-AGGREFACT leaderboard (Tang et al., 2024) exceeding a BACC of $\approx 61\%$ (Appendix F, Table 8). The primary denominator is therefore the *core grounded* four –MEDHALLU, RAGTRUTH, XSUM, and WICE– with all six grounded sets as a secondary denominator and the suite of eight as the widest.

Over those benchmarks, we set $N_v = 10$, instantiating the panel with ten open-weight instruction-tuned judges of 4–9B parameters drawn from disjoint model families to diversify failure modes: `Llama`, `Qwen`, `Mistral`, `Gemma`, `Phi`, `Yi`, `Command-R`, `GLM`, `Granite`, and `Falcon` being the family tag by which we refer to each throughout (for details see Appendix K, Table 21). A member returning no parseable verdict is recorded as abstaining. The prompt supplies the task definition, and takes one of two frames: the rubric we label `def_on` on the four question-answering sets, scaffolding the tuple (question, evidence, candidate) on MEDHALLU and RAGTRUTH and dropping the evidence block on the ungrounded FACTSCORE and TRUTHFULQA; and a MiniCheck-style (Tang et al., 2024) *claim-support* frame, pairing (document, claim), on the four LLM-AGGREFACT claim-verification sets (WICE, XSUM, CNN, EXPERTQA), whose structure requires it (Appendix F). The frame is fixed per dataset and identical across all ten members and the frontier judge.

[2] On MEDHALLU the stacked panel spans $[0.695, 0.756]$ and its best single judge $[0.658, 0.727]$, largely overlapping, while their paired difference is $+0.032$ $[+0.004, +0.060]$.

[3] Question-answering grounded: MEDHALLU and RAGTRUTH; claim-verification grounded: WICE, XSUM, CNN, and EXPERTQA (from LLM-AGGREFACT); ungrounded: FACTSCORE and TRUTHFULQA.

Table 1: Comparator ladder over the eight benchmarks. Columns are ordered by the labelled information each comparator consumes. κ is given as a decimal and BACC as a percentage.

		label-free			*single judge, label-selected*			*stacked panel*		*frontier* (SONNET)		*paired Δ (panel − comparator)*			
Dataset	n	avg. mem.	unw.	D–S	best (avg)	best (CV)	oracle	κ	BACC	κ	BACC	avg (κ)	best (κ)	SONNET (κ)	SONNET (BACC)
Core grounded															
MEDHALLU	2000	0.590	0.708	0.716	0.656	0.693†	0.693	0.725	86.3	0.761	88.1	**+0.135**	**+0.032**‡	**−0.036**	**−1.79**
WICE	222	0.385	0.449	0.486	0.547	0.484†	0.547	0.540	77.0	0.548	77.4	**+0.155**	+0.056	−0.008	−0.39
RAGTRUTH	1362	0.181	0.179	0.325	0.399	0.399	0.399	0.409	70.4	0.601	80.0	**+0.228**	+0.010	**−0.192**	**−9.62**
XSUM	546	0.312	0.395	0.420	0.377	0.400†	0.437	0.437	71.9	0.418	70.9	**+0.125**	+0.037	+0.019	+0.96
Grounded (secondary)															
CNN	114	0.088	0.070	0.088	0.298	0.298	0.298	0.349	67.5	0.469	73.5	**+0.261**	+0.051	−0.120	−6.05
EXPERTQA	1462	0.154	0.146	0.205	0.217	0.240	0.240	0.235	61.8	0.165	58.2	**+0.082**	−0.005	**+0.070**	**+3.50**
Ungrounded (contrast)															
FACTSCORE	330	0.259	0.485	0.515	0.444	0.444†	0.444	0.503	75.2	0.552	77.6	**+0.244**	+0.059	−0.049	−2.45
TRUTHFULQA	1634	0.274	0.374	0.409	0.471	0.471†	0.471	0.472	73.6	0.623	81.1	**+0.198**	+0.000	**−0.151**	**−7.56**
count ↑/↓												**8**(–)/–(–)	**1**(6)/–(1)	**1**(1)/**3**(3)	**1**(1)/**3**(3)

avg. mem.: average member; *unw.*: unweighted majority; *D–S*: Dawid–Skene. The four rightmost columns are paired cluster-bootstrap differences on the same resampled clusters ($B = 2000$), **bold** where the 95% interval excludes zero (cf. Figure 3(b)); the *count* line tallies them as improvement (↑) or worsening (↓), as '**clear** (unresolved)', read before rounding. ‡: beats the oracle too; underline: the frontier is not clearly ahead; †: an under-covering CV-best baseline (Appendix K).

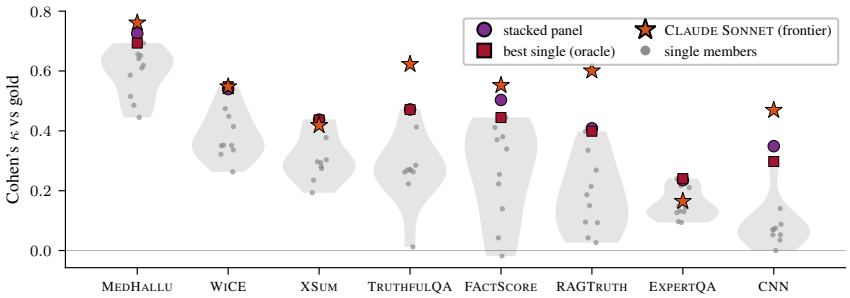


Figure 2: Single-member competence per dataset (Cohen's κ vs gold). Violins and strips carry the ten judges' κ, with the stacked panel (●), the oracle best single (■), and the frontier judge SONNET (⋆; §4) overlaid. Datasets are ordered by descending mean single-member competence.

3 CORE RESULTS

We report the study's core results in three steps: first the panel measured against the other ways of combining the same ten votes, then against the best single judge *inside* it, and finally the member-level account of when either contrast comes out in the panel's favour (§3.3); the comparison with the frontier judge is deferred to §4. Table 1 carries all of it on one comparator ladder, whose columns are ordered by the labelled information each baseline consumes: from the label-free aggregators a deployer can always run, through the one globally best member, to the per-dataset best single and the frontier judge. Reading a row left to right is therefore reading what each increment of supervision buys. The full per-method ladders and the rescue and aggregator contrasts are in Appendix I (Table 17).

3.1 THE PANEL AGAINST ITS OWN NAIVE VOTE

Which panel is on offer must be established first, and the answer is not the obvious one: combining the ten votes naively gives a panel that has to be rescued from itself. Wherever competence is concentrated in one leader or spread unevenly across a field with actively harmful members, the unweighted majority panel sits *significantly below* the best single judge –by 0.094 to 0.228 κ on RAGTRUTH, TRUTHFULQA, CNN, and EXPERTQA– because the weaker, mutually redundant members outvote the few capable ones. Stacking reverses this: reading the *unw.* and *stacked* columns of Table 1 against *best (CV)*, the panel is no longer clearly below the best single judge on any of the four, and the paired delta (stacked − unweighted) is significantly positive on each (up to +0.279 on CNN) and borderline on MEDHALLU (intervals in Appendix I, Table 17). The effect is, by contrast, small or absent where the naive vote is not clearly below the best single judge, i.e. where competence is spread (MEDHALLU, WICE, FACTSCORE) or redundant (XSUM); the size of the rescue thus depends on how heterogeneous the members are. Weighting each vote by that member's

own agreement with gold (each in isolation) gives the graded form of the same point: both accuracy- and log-odds-weighted voting sit below the stacker on every dataset (Appendix I).

Comparison with an unsupervised aggregator. A sharper test than the naive vote is another *aggregator* over the same ten votes, and the natural choice is the model of Dawid & Skene (1979) (*D–S*), which estimates each member's reliability without gold and so isolates what the labels buy (Appendix L). The stacked panel is never significantly beaten by it (Appendix I), and the delta (stacked − DS) offers a clear improvement on RAGTRUTH, TRUTHFULQA, and CNN, and straddles zero elsewhere; D–S itself falls significantly below the CV-best single wherever the best member must be picked out of a weak or correlated field. This is the competence/correlation account (§3.3 below) seen from the aggregator side: where errors decorrelate and competence is spread, an unsupervised aggregator suffices; if one judge must be recovered, only supervised re-weighting does.

3.2 THE PANEL AGAINST THE BEST SINGLE JUDGE

Against the fair (CV-best) baseline, the paired difference is an improvement on all four core grounded datasets, and on one of them it is clear: MEDHALLU offers a delta (stacked − best single) $= +0.032$ clear against both the CV-best and the oracle (`Granite`) –the suite's one *clean win* in the sense of §2– and still clear on the items the baseline commits on ($+0.031$ $[+0.003, +0.058]$); on the other three, the improvement stands but the sample does not resolve it. A Holm correction (Holm, 1979) over the four core deltas (cf. §2) leaves that win significant (adjusted one-sided $p = 0.048$), whereas over all eight it would not ($p = 0.096$, see also Appendix K). With nine competent members under the claim-support frame ($\kappa \geq 0.30$; cf. Figure 3(a)) and the best single being already strong, WICE offers an unresolved improvement ($+0.056$) at $n = 222$. On RAGTRUTH, conversely, one member (`Gemma`) dominates a weak field and is reliably found by cross-validation, leaving almost no headroom ($+0.010$). Finally, on XSUM the members carry moderate, redundant competence (four at $\kappa \geq 0.30$) and the edge ($+0.037$, one-sided $p = 0.045$) is the suite's closest unresolved improvement. The two grounded sets held back are unresolved as well (CNN as an improvement and EXPERTQA as a worsening), and are reported in Appendix I. This fixes the scope of the internal contest: the difference favours the panel on seven out of eight datasets, yet is clear on one core set of four.

Comparing baselines needing no per-dataset labels. Identifying the best single, however, requires labels a deployer might not have; against the baselines needing no such per-dataset selection (Table 1, leftmost block) the panel is strong wherever there is signal: it improves on the label-free average member by $+0.082$ to $+0.261$ κ, clearly on every dataset in the suite, and is nowhere clearly below the single best-on-average judge (`Gemma`). The one-of-four verdict above is thus the same panel read against a baseline that has already spent those labels, and its position on that ladder is what §4 prices.

3.3 THE COMPETENCE/CORRELATION ACCOUNT

The pattern of paired differences above is not arbitrary: it is ordered by two member-level quantities recovered from the calibration set the aggregator already requires, namely the *distribution of competence* –how many members are individually capable, rather than how good the best one is– and the *correlation of errors* — whether the members, when wrong, are wrong on the same items, which we summarise by their mean pairwise error-correlation $\bar{\phi}$. The ordering, however, does not rest on eight dataset points alone: restricting to the items on which the best single judge errs isolates the recovery the *other* members carry. Under this framing, the panel recovers the item more often precisely when a larger fraction of members is individually correct on it (logistic regression with dataset fixed effects, coefficient $+4.67$; point-biserial $r = +0.41$, $p < 10^{-20}$, positive within all eight datasets; Appendix B). Distributed, decorrelated competence is thus what the panel harnesses, and what the stacker's own weights reward in place of average accuracy (Appendix I, Figure 10).

The two coordinates order the suite, and Figure 3(a) places the eight datasets in the plane they span (qualitatively, eight points being too few to resolve either coordinate statistically, cf. Appendix B, Table 5). The panel's improvement over the best single is largest where competence is distributed over several capable members whose errors decorrelate, shrinks where that competence is moderate and mutually redundant, and vanishes where one model dominates a weak field. The same ordering governs the frontier comparison. Read against all three comparators at once (Figure 3(b)), the difference over the *best* single narrows to zero as one model comes to dominate, and the difference against the *frontier* judge worsens along with it (§4 below), while the difference over the *average*

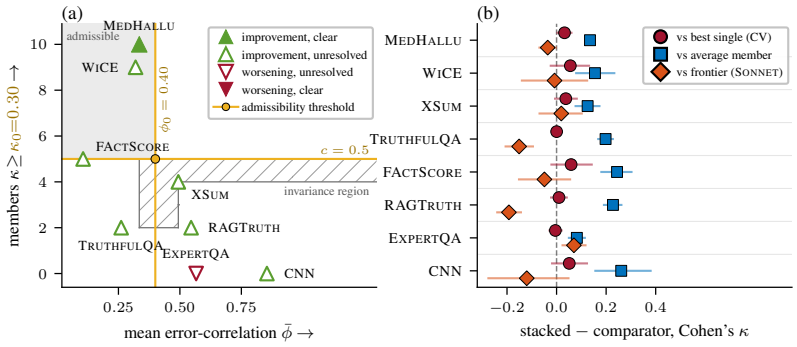


Figure 3: The regime ordering. **(a)** The eight datasets (markers) scattered by the members' mean pairwise error-correlation $\bar{\phi}$ (x-axis) against the number of members at $\kappa \geq \kappa_0 = 0.30$ (y-axis). The gold lines are the thresholds $\phi_0 = 0.40$ and $c = 1/2$ (five members), and bound the admissible region (shaded quadrant); the hatched area is the invariance region, the positions of the gold point at which no dataset changes side, unbounded to the right. Varying κ_0 moves the points rather than an edge (Appendix C). **(b)** The panel's paired delta against the CV-best single (●), the average member (■), and the frontier judge (◆), reported per dataset with 95% paired cluster-bootstrap intervals.

member does neither, staying uniformly large wherever there is signal. The regime in which a panel improves on its own leader is thus also one in which the frontier is not materially ahead. The frontier has to be kept only in the dominant-leader regime the test rejects, which is why we answer the two questions together. Interestingly, it is not the regime the literature would predict: grounded-versus-ungrounded is numerically present but mechanistically empty, the two ungrounded sets falling in *opposite* regimes, and what moves with the task is instead the identity of the best single judge — which is what a panel is meant to supply, and what fails to transfer (Appendices B and E).

The admissibility test. The account enters the method only if a deployer can tell, on a new task, which side of it they are on: this is the second component of RAIM (§1), which, along with the stacker's down-weighting of redundant members, makes the scheme *robust*. The test reads the two coordinates off the calibration set, and admits a task when enough members are competent and their errors are weakly correlated. Formally, given a competence bar κ_0, a share c, and a correlation bar ϕ_0, a task is *admissible* when at least a share c of the members reach $\kappa \geq \kappa_0$ and $\bar{\phi} \leq \phi_0$; in the plane of Figure 3(a), at fixed κ_0, this is a quadrant, and, in what follows, we adopt $(\kappa_0, c, \phi_0) = (0.30, \frac{1}{2}, 0.40)$ as an operating point, resulting in MEDHALLU, WICE, and FACTSCORE being admitted. Tightening any of the three admits fewer tasks, and where to stop depends on what a deployer loses by dropping a frontier judge that was needed, against paying for one that was not; moreover, with just eight datasets, the calls change only where a threshold crosses a dataset's own coordinate (hatched region of Figure 3(a); Appendix C). What holds at every operating point is the direction of the test: over a grid of shares and correlation bars, and on 1,888 (sub-)panels drawn from the same ten members, the admitted panels are more often within reach of the frontier judge, and more often above their own best member, than the rejected ones (Appendix C). Therefore, stackers on admissible tasks are expected to improve on the CV-best member; no task failing the test shows a clear gain over that member, and on dominant-leader ones, the frontier stays materially ahead (§4).

The two coordinates are not, however, equally costly. The correlation one is free: two binary verdicts against a common gold label agree exactly when both are equally right or wrong, so the members' mean pairwise *agreement* equals that of their error indicators and needs no gold (only its normalisation into $\bar{\phi}$ does), and it orders the suite like $\bar{\phi}$ (Spearman $\rho = +0.93$). The competence one has no assumption-free unsupervised counterpart: a Dawid–Skene fit to the same votes (§3.1) rates all ten members competent where, e.g., on CNN and EXPERTQA none clears $\kappa = 0.30$. On calibration samples, the regime call is reproduced 85.5% of the time from fifty labelled records and 91.9% from a hundred, the budgets at which the aggregator itself becomes worth fitting, so the diagnostic and the calibration are one purchase, whereas naming the best member from a hundred succeeds only 69% of the time (Appendix C). Notably, the error-correlation is largely inherited from pretraining rather than induced by alignment (Appendix D), and on four of the eight datasets our cheap panel still carries more independent votes than Kohli's (2026) nine frontier judges reach on any of theirs (Appendix B).

Table 2: Quality, cost and applicability of each route. BACC is evaluated over three nested groups: the four LLM-AGGREFACT sets, the six grounded sets and the full suite. Published cells (top) are from the literature (Tables 8 and 9; κ in Table 12).

		quality (BACC) *by scope*			*cost* (USD)		*scope*	
Route	size	LLM-AGGREFACT-4	grounded 6	suite 8	per 1k	one-time	open	ungr.
purpose-trained and prompted comparators								
`Bespoke-MiniCheck-7B` (leaderboard)	7B	71.4	—	*n/a*	0.004¶	fine-tune	●	○
`MiniCheck-Flan-T5-L` (Tang et al., 2024)	0.8B	68.8 (68.4†)	66.6†	*n/a*	0.004¶	fine-tune	●	○
`Granite Guardian 3.3` (leaderboard)	8B	69.5	—	—	0.004¶	fine-tune	●	⊙
`Prometheus-2` (Kim et al., 2024b)	7B	59.3†	60.0†	60.4†	0.004¶	fine-tune	●	●
`GPT-4o`, prompted (leaderboard)	—	70.8	—	—	1.93	*none*	○	●
measured here, one protocol across all eight sets								
average cheap open judge	4–9B	59.6	62.4	62.5	0.004	*none*	●	●
best single member (CV-selected)	4–9B	67.7	68.5	69.3	0.004	$100–$200	●	●
larger open judge (`Qwen2.5-32B`)	32B	69.4	72.1	72.6	0.008	*none*	●	●
stacked cheap panel (this work)	10 × 4–9B	**69.5**	**72.5**	**73.0**	0.039	$100–$200	●	●
frontier API judge (SONNET)	—	70.0	74.7	75.9	2.52	*none*	○	●

†: our rerun under the system's released weights; —: no data; *n/a*: scope the judge cannot be run on; *ungr.*: applicability to ungrounded factuality (● yes, ⊙ partly, ○ no); *cost*: inference per 1,000 items (API or cloud GPU rates, Appendix H) and one-time outlay on a new domain ($1 per label); ¶: estimated by comparison.

4 STANDING IN FOR A FRONTIER JUDGE

We now ask what is lost and what is gained by running the panel instead of paying for a frontier judge. We adopt CLAUDE SONNET (Anthropic, 2025) (`claude-sonnet-4-6`; temperature 0) as frontier judge, and difference it against the panel item by item under the same paired cluster bootstrap (§2).

The quality cost. Read as eight paired differences (the two rightmost columns of Table 1; intervals in Appendix H, Table 13), the substitution offers an improvement on two datasets (one of which is *clear*) and worsens on the remaining six (three *unresolved* and three clear, one of which is negligible). The clear improvement is EXPERTQA ($\Delta\kappa = +0.070$ $[+0.024, +0.117]$), the set on which SONNET is weakest by far ($\kappa = 0.165$, the next-weakest being 0.418), below even the panel's best single member (Figure 2), a shortfall not ascribable to the prompt frame, the better of the two we ran for SONNET (Appendix F). XSUM's improvement and WICE's, FACTSCORE's, and CNN's worsenings are all unresolved, leaving those four statistically too close to the frontier to claim a result in either direction. Of the three clear worsenings, MEDHALLU's is too small to matter in deployment ($\Delta\kappa = -0.036$ $[-0.069, -0.003]$, i.e. 1.8 points of balanced accuracy on a task the panel already scores at 86.3), leaving only RAGTRUTH (-0.192) and TRUTHFULQA (-0.151) materially adverse: precisely the two on which one strong judge already captures the task, so the panel has nothing worth aggregating and a frontier model is simply a better single judge. The results are comparable when read on BACC (which charges every abstention on the full denominator): the panel gives up 2.9 points on average (73.0 against 75.9) and at most 9.6 (RAGTRUTH), while gaining 3.5 on EXPERTQA and 1.0 on XSUM. A frontier judge therefore has to be kept in one regime only: the panel holds a median 93% of the frontier's κ and is clearly and materially behind on two of eight, and the account of §3.3 names those two as the dominant-leader sets — a reading available from the members themselves, at the same labelling cost the aggregator already requires (Appendix C).

Cost and deployment profile. Table 2 prices that deficit against what a deployer would weigh. The ten members together run at $\$0.039$ per thousand items against $\$2.52$ for the SONNET API, about 1.55% of the price (Appendix H derives both; Appendix M bounds them). Moreover, the panel runs on premises, so no evaluated text leaves the deployer's infrastructure and no rate limit or model deprecation enters the evaluation loop, a solution the escalation route cannot offer (Appendix L). Inference is not, however, where the panel's money goes. The stacker needs in-domain labels, so the break-even is a volume rather than a rate, and how large a volume turns on how many labels the aggregator requires, which we measure at fifty to a hundred labelled records (Appendix C, Figure 6(b)), the same operating range as the diagnostic of §3.3. Priced at $1 per label, and two labels to a record on the paired sets, that is an outlay of $100 to $200, which the panel's cheaper inference repays between some $40{,}000$ and $81{,}000$ evaluated items — a large yet sustainable volume.

Comparison with the LLM-AGGREFACT leaderboard. Four of our eight sets are subsampled from the LLM-AGGREFACT test split and read in balanced accuracy, the metric that leaderboard reports; Appendix F sets out what licenses reading our numbers beside the published ones, showing an

agreement within about one point of balanced accuracy, so gaps under three points are not informative. On those four, the panel reaches 69.5, drawing level with `Granite Guardian 3.3` and sitting inside the agreement band of the leaderboard's top system (`Bespoke-MiniCheck-7B`, 71.4) and of `GPT-4o` (70.8); all this from a field whose best member scores 67.7 (average member 59.6). This is, however, the comparison least favourable to the panel, with specialists being scored on the datasets they were built for: run at their documented operating points across our eight datasets, the ordering materially reverses (Table 2; per-dataset in Appendix G); and the strongest specialist we reproduce, `MiniCheck-Flan-T5-L`, scores 66.6 against the panel's 72.5 across the six grounded sets, 63.1 against 86.3 on MEDHALLU alone, and cannot be evaluated on the full suite as its interface requires a document the two ungrounded sets cannot supply. The evaluation-tuned judges, which do run everywhere, are weak throughout (`Prometheus-2` at 60.4 over the suite, against our 73.0). A cheap panel does not beat a trained detector on the detector's own ground, the two serving different ends: a specialist gives the best quality on the distribution it was trained for, while a generalist panel gives comparable verdicts that keep working if the domain moves and the reference disappears.

Outside LLM-AGGREFACT. Two further benchmarks admit a published comparison once the metric is matched, MEDHALLU and RAGTRUTH being reported in hallucinated-class F1 (Appendix F, Table 9). On MEDHALLU the stacked panel reaches 86.2, drawing level with the strongest published open judge (`Qwen2.5-14B`, 85.2) and with `GPT-4o-mini` (84.1), while resting within two points of the published `GPT-4o` (87.7). Three of the published detectors are *members of our own panel*, which gives a second harness calibration (Appendix F). On RAGTRUTH, however, the published numbers sit on that corpus's test split, at a hallucinated rate of 0.178 against the 0.5 our construction imposes, and hallucinated-class F1 moves with prevalence, so our rows are not comparable in level with them. The anchors themselves show the pattern of this paper, with the strongest prompt-only system (`GPT-4-turbo`, 45.6) falling far below the three detectors fine-tuned on the corpus (68.2 to 74.8). On MEDHALLU prevalence is matched; however, we charge abstentions as missed detections and the two sides are not scored on the same items, so results compare levels and license no ranking.

Open-judge scaling. Size alone does not close the remaining gap either: a larger *open* judge does not improve monotonically with scale, `Qwen2.5-32B` being the better of the two on four sets, and `Qwen2.5-72B` trails SONNET on four of the eight, drawing level mostly where the frontier is weakest (Appendix H). Within the one open family we scaled, parameter count is therefore not what separates an open judge from a frontier model. A single `Qwen2.5-32B` is nonetheless a close rival on average (72.6 against 73.0 suite BACC, Table 2), yet nothing tells a deployer in advance where it will hold, whereas the panel's regime is read from its own members.

5 CONCLUSION

In this work, we studied whether a panel of cheap open-weight judges can stand in for a frontier one, what the substitution sacrifices, and when it is worth making. We built that panel from ten 4–9B members of uneven and, on any new task, unknown competence: none of them clears our competence bar on all eight benchmarks, and naming the strongest from a hundred labelled records succeeds only 69% of the time. Without knowing which to trust, we still read from all ten a verdict at least as good as the best of them, holding a median 93% of CLAUDE SONNET's κ and giving up 2.9 points of balanced accuracy on average, for a sixty-fourth of the inference cost, and with nothing leaving the deployer's premises. Out of the eight analysed benchmarks, the panel clearly leads on one and falls materially behind on only two. The panel's cost is a one-time in-domain calibration of 50–100 records, amortised against SONNET after some 40–81K items. Against the label-free ways of reading the same votes the gain is unconditional, the panel being clearly ahead of the average member everywhere and never clearly beaten by an unsupervised aggregator; against its own best member it is not, the panel leading where competence is spread and errors decorrelate, and otherwise recovering its leader. Both coordinates of that condition are read from the same calibration set the aggregator already consumes, so a deployer can estimate the regime before committing to the panel. The main limitation is that the panel is capped by its roster: the gain requires several individually capable members whose errors decorrelate, a condition that might drift from roster to roster and from one model version to the next, and one we characterise for a single fixed roster of frozen generalists. Adding members pays only where they already decorrelate from one another: since error-correlation rises with model accuracy (Kim et al., 2025; Goel et al., 2025), larger members should not be expected to relax that condition, which makes roster composition the lever worth pulling.

REPRODUCIBILITY STATEMENT

All code, together with the per-dataset run artefacts from which every result is produced, will be made fully available upon acceptance as two GitHub repositories, split where a reader could regenerate everything without intensive or costly computation.[4] `raim-verdicts` carries the full inference pipeline, from dataset construction and prompt rendering to the model backends and the verdict parser, and is readily extensible to further models (mainly via Hugging Face) or benchmarks (mainly via the Python `datasets` library); `raim-analysis` carries everything that regenerates from the verdicts, namely the stacked aggregator and cross-fit, the cluster-bootstrap machinery, and the table- and figure-generation pipeline. Full verdicts, including paid API generation, will be released as well in a public archive.

Both repositories will be released under CC BY-NC-SA 4.0. That licence cannot cover the benchmarks, whose terms a `NOTICE` file records source by source; we distribute only our own verdict records, keyed to a content-addressed index against which a third party can verify the join to each publisher's data without our redistributing any source text. The decision rule and the bootstrap protocol are specified in §2, the judge prompt, decoding constraints, and coverage accounting in Appendix K, and the full per-dataset ladders in Appendix I.

Every reported number other than the re-run agreements below is generated programmatically from in-repo artefacts: the `QUICKSTART` of `raim-analysis` gives the commands that verify the data, rebuild every table and figure from the committed derived tier in minutes, print the headline numbers of the text beside the files they are read from, and regenerate that tier from the plain verdicts within about an hour, with no GPU and no network, under Python 3.12 to 3.14 on macOS and Linux. The analysis is deterministic at `seed` $= 0$ on a given machine, whereas across hardware the order in which the numerical libraries accumulate sums perturbs the ninth decimal and beyond; the release therefore checks a regenerated tier against the committed one at a 10^{-8} tolerance on floating-point values and exactly on counts, labels, judging conditions, and panel membership, so that a number which has really moved fails the check whereas a change of hardware does not. Inference, conversely, is not bit-reproducible across GPU model or serving stack, which is why the verdicts are released as data: a full re-run on independent hardware, with vLLM and `transformers` a major version ahead, moved individual verdicts (median per-judge agreement 0.98 for the generative panel, 0.99 for the constrained scorer) but left every win/tie verdict of §3 unchanged.

AI USE STATEMENT

Generative AI enters this work as part of the study, mainly through the generation of the data (the judgements) we analyse. The items being judged come from the literature: they are the eight released benchmarks' own instances, collected by their authors –some of whose candidates were themselves produced by language models when those benchmarks were built– and used here unmodified, so no synthetic dataset was created for this study. What our pipeline does generate is the evidence the paper rests on, i.e. the per-item verdicts of the ten open-weight judges and of the frontier judge, and generative AI assisted in writing the code that elicits them: the model backends, the prompt construction, and the verdict parser. Thus, our verdict set (the released artefacts) collects the outputs of generative AI models, and the paper's results are derived from them, yet the AI itself does not generate the artefact, nor does it make any of the claims or conclusions we draw from it.

We disclose AI support under *implement methods* and, equivalently, under *create and edit software code*, the two being one activity here; the main assistance concerns the source code that calls the LLMs in the panel, and extends to parts of the aggregation and cluster-bootstrap machinery and of the table- and figure-generation pipeline, as well as to the preparation of a submission-grade artefact. We further disclose AI support under *provide feedback on research methodology or experiments*, since automated AI review tools critiqued the experimental design and the analysis (including the PAT automated pre-submission feedback offered on OpenReview by ICLR Program Chairs); each suggestion was analysed and verified against the code and the results, prior to any change by the authors.

[4] An anonymised snapshot of both repositories accompanies this submission as supplementary material.

We did not use generative AI tools to develop the conceptual framework, to formulate or prove mathematical claims, to propose or refine the hypotheses, to design the experimental methodology and the decision rule, to interpret the results, or to translate any part of this work, all of which are the authors' own; the remaining tasks requiring disclosure are not applicable here.

We additionally used generative AI tools to draft and edit parts of the manuscript for readability and to identify relevant literature. We have reviewed all AI-assisted work. Every result of ours is regenerated programmatically from the released run artefacts and was verified against them, except for the agreements reported in the Reproducibility statement for the independent re-run, which compare our verdicts with those of a second run that, for reasons of space, we report only in summary. Every published value set beside ours was checked against its source, all AI-assisted code was read and extensively tested by the authors, and every reference so identified was retrieved and read in the original. We take responsibility for the final content of this work, including text, claims, and artefacts produced with the aid of generative AI.

ETHICS STATEMENT

This work evaluates existing model outputs and collects no new human-generated data: all eight benchmarks are publicly released datasets with published gold labels, used under their respective licences and terms, and no human subjects or annotators were involved. We redistribute only derived per-item judge verdicts and aggregate results, never the source datasets, which can be fetched from their publishers (all but FACTSCORE through the Python `datasets` module). The method's purpose, i.e. detecting unsupported content in generated text, is safety-positive and, being an evaluator rather than a generator, offers little scope for misuse. One benchmark (MEDHALLU) contains medical question-answering material with synthesised hallucinations; our detector is a research artefact evaluated on that benchmark's terms and is not a clinical tool.

REFERENCES

01.AI. Yi: Open foundation models by 01.AI, 2024. ARXIV:2403.04652.

Junade Ali. Quantifying diversity of thought: A predictive law of weighted LLM ensemble lift, 2026. ARXIV:2607.17384.

Anthropic. Claude Sonnet. `https://www.anthropic.com/claude`, 2025.

Krishnakumar Balasubramanian, Aleksandr Podkopaev, and Shiva Prasad Kasiviswanathan. Dependence-aware label aggregation for LLM-as-a-judge via Ising models, 2026. ARXIV:2601.22336.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. HalluLens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 24128–24156, 2025. doi: 10.18653/v1/2025.acl-long.1176.

Forrest Sheng Bao, Miaoran Li, Renyi Qu, Ge Luo, Erana Wan, Yujia Tang, Weisi Fan, Manveer Singh Tamber, Suleman Kazi, Vivek Sourabh, Mike Qi, Ruixuan Tu, Chenyu Xu, Matthew Gonzales, Ofer Mendelevitch, and Amin Ahmad. FaithBench: A diverse hallucination benchmark for summarization by modern LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL), Volume 2: Short Papers*, pp. 448–461, 2025. doi: 10.18653/v1/2025.naacl-short.38.

Loris Bergeron, Ioana Buhnila, Jérôme François, and Radu State. HalluGuard: Evidence-grounded small reasoning models to mitigate hallucinations in retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 16918–16932, 2026. doi: 10.18653/v1/2026.findings-acl.835.

Hongyu Chen and Seraphina Goldfarb-Tarrant. Safer or luckier? LLMs as safety evaluators are not robust to artifacts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 19750–19766, 2025. doi: 10.18653/v1/2025.acl-long.970.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024. ARXIV:2305.05176.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Cohere. Command R7B. `https://cohere.com/blog/command-r7b`, 2024.

Manuel Cossio. A comprehensive taxonomy of hallucinations in large language models, 2025. ARXIV:2508.01781.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979. doi: 10.2307/2346806.

Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris, 1785.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems (MCS 2000)*, volume 1857 of *Lecture Notes in Computer Science*, pp. 1–15. Springer, 2000. doi: 10.1007/3-540-45014-9_1.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Gemma Team, Google DeepMind. Gemma 2: Improving open language models at a practical size, 2024. ARXIV:2408.00118.

GLM Team. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools, 2024. ARXIV:2406.12793.

Ameya Godbole and Robin Jia. Verify with caution: The pitfalls of relying on imperfect factuality metrics. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22889–22912, 2025. doi: 10.18653/v1/2025.findings-acl.1175.

Shashwat Goel, Joschka Strüber, Ilze Amanda Auzina, Karuna K. Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines AI oversight. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025. ARXIV:2502.04313.

Granite Team, IBM. Granite 3.0 language models. `https://github.com/ibm-granite/granite-3.0-language-models`, 2024.

Benjamin Grayzel. Cost-effective automated judging of natural-language mathematical proofs, 2026. ARXIV:2608.00004.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. A survey on LLM-as-a-judge. *The Innovation*, 7(6):101253, 2026. doi: 10.1016/j.xinn.2025.101253.

Steve Han, Gilberto Titericz Junior, Tom Balough, and Wenfei Zhou. Judge's verdict: A comprehensive analysis of LLM judge capability through human agreement, 2025. ARXIV:2510.09738.

David Hartmann, Manuel Tonneau, Angelie Kraft, LK Seiling, Dimitri Staufer, Pieter Delobelle, Jan Fillies, Anna Ricarda Luther, Jan Batzner, and Mareike Lisker. Bye bye Perspective API: Lessons for building and governing measurement infrastructure, 2026. ARXIV:2604.25580.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Elias Hossain, Niloofar Yousefi, and Ser-Nam Lim. Agreement overstates evidence: Error dependence in LLM judge consensus, 9 2026. ARXIV:2609.22512.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1120–1130, 2013. URL https://aclanthology.org/N13-1132/.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5880–5895, 2025. doi: 10.18653/v1/2025.findings-acl.306.

Ariel Jaffe, Ethan Fetaya, Boaz Nadler, Tingting Jiang, and Yuval Kluger. Unsupervised ensemble learning with dependent classifiers. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51 of *Proceedings of Machine Learning Research*. PMLR, 2016. URL https://proceedings.mlr.press/v51/jaffe16.html. ARXIV:1510.05830.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023. ARXIV:2310.06825.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. URL https://neurips.cc/virtual/2025/loc/san-diego/poster/121421.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *International Conference on Learning Representations (ICLR)*, 2025. ARXIV:2407.18370.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7561–7583, 2023. doi: 10.18653/v1/2023.emnlp-main.470.

Donghwan Kim. Are diversity metrics measuring diversity? A capability-controlled audit of majority-vote gain in LLM ensembles, 2026. ARXIV:2607.20768.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025. ARXIV:2506.07962.

Hyun-Chul Kim and Zoubin Ghahramani. Bayesian classifier combination. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22 of *Proceedings of Machine Learning Research*, pp. 619–627. PMLR, 2012. URL https://proceedings.mlr.press/v22/kim12.html.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations (ICLR)*, 2024a. ARXIV:2310.08491.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4334–4353, 2024b. doi: 10.18653/v1/2024.emnlp-main.248.

Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels, 2026. ARXIV:2605.29800.

Ádám Kovács and Gábor Recski. LettuceDetect: A hallucination detection framework for RAG applications, 2025. arXiv:2502.17125.

Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003. doi: 10.1023/A:1022859003006.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453.

Anish Laddha, Nitesh Pradhan, and Gaurav Srivastava. SLMJury: Can small language models judge as well as large ones?, 2026. arXiv:2606.07810.

Deren Lei, Yaxi Li, Siyao Li, Mengya Hu, Rui Xu, Ken Archer, Mingyu Wang, Emily Ching, and Alex Deng. FactCG: Enhancing fact checkers with graph-based multi-hop data. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 5002–5020, 2025. doi: 10.18653/v1/2025.naacl-long.258.

Haifeng Li and Mo Hai. State-dependent error correlations shape voting thresholds in committees of AI agents, 2026. arXiv:2607.23931.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. LLMs-as-judges: A comprehensive survey on LLM-based evaluation methods, 2024a. arXiv:2412.05579.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. Generative judge for evaluating alignment. In *International Conference on Learning Representations (ICLR)*, 2024b. arXiv:2310.05470.

Yanran Li. Calibrate, don't curate: Label-efficient estimation from noisy LLM judges, 2026. arXiv:2605.09702.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229.

Llama Team, AI@Meta. The Llama 3 herd of models, 2024. arXiv:2407.21783.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 3025–3045, 2024. doi: 10.18653/v1/2024.naacl-long.167.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9004–9017, 2023. doi: 10.18653/v1/2023.emnlp-main.557.

Microsoft. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs, 2025. arXiv:2503.01743.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12076–12100, 2023. doi: 10.18653/v1/2023.emnlp-main.741.

Yasmin Moslem and John D. Kelleher. Dynamic model routing and cascading for efficient LLM inference: A survey. *Transactions on Machine Learning Research*, 2026. arXiv:2603.04445.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1797–1807, 2018. doi: 10.18653/v1/D18-1206.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 10862–10878, 2024. doi: 10.18653/v1/2024.acl-long.585.

Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. MedHallu: A comprehensive benchmark for detecting medical hallucinations in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2858–2873, 2025. doi: 10.18653/v1/2025.emnlp-main.143.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Emmanouil Antonios Platanios, Avrim Blum, and Tom M. Mitchell. Estimating accuracy from unlabeled data. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014. URL `https://dl.acm.org/doi/10.5555/3020751.3020822`.

Qwen Team. Qwen2.5 technical report, 2024. ARXIV:2412.15115.

Hossein A. Rahmani, Emine Yilmaz, Nick Craswell, and Bhaskar Mitra. JudgeBlender: Ensembling automatic relevance judgments. In *Companion Proceedings of the ACM Web Conference 2025 (WWW)*, pp. 1268–1272, 2025. doi: 10.1145/3701716.3715536.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017. doi: 10.14778/3157794.3157797.

Emanuele Ricco, Elia Onofri, Lorenzo Cima, Stefano Cresci, and Roberto Di Pietro. A geometric analysis of small-sized language model hallucinations. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. ARXIV:2602.14778.

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 (ICML)*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025. ARXIV:2501.17178.

Wooseok Seo, Seungju Han, Jaehun Jung, Benjamin Newman, Seungwon Lim, Seungbeen Lee, Ximing Lu, Yejin Choi, and Youngjae Yu. Verifying the verifiers: Unveiling pitfalls and potentials in fact verifiers. In *Proceedings of the Second Conference on Language Modeling (COLM)*, 2025. ARXIV:2506.13342.

Akshita Singh, Prabesh Paudel, and Siddhartha Roy. Hallucination detection via activations of open-weight proxy analyzers, 2026. ARXIV:2605.07209.

Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, and Cheng Niu. RAG-HAT: A hallucination-aware tuning pipeline for LLM in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1548–1558, 2024. doi: 10.18653/v1/2024.emnlp-industry.113.

Guangzhi Sun, Anmol Kagrecha, Potsawee Manakul, Phil Woodland, and Mark Gales. SkillAggregation: Reference-free LLM-dependent aggregation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 15532–15548, 2025. doi: 10.18653/v1/2025.acl-long.756.

Manveer Singh Tamber, Forrest Sheng Bao, Chenyu Xu, Ge Luo, Suleman Kazi, Minseok Bae, Miaoran Li, Ofer Mendelevitch, Renyi Qu, and Jimmy Lin. Benchmarking LLM faithfulness in RAG with evolving leaderboards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track (EMNLP)*, pp. 799–811, 2025. doi: 10.18653/v1/2025.emnlp-industry.54.

Liyan Tang, Tanya Goyal, Alexander R. Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin F. Rousseau, and Greg Durrett. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 11626–11644, 2023. doi: 10.18653/v1/2023.acl-long.650.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8818–8847, 2024. doi: 10.18653/v1/2024.emnlp-main.499.

Technology Innovation Institute. The Falcon 3 family of open models. https://huggingface.co/blog/falcon3, 2024.

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, 2024. arXiv:2404.18796.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.05087.

Jacob Whitehill, Paul Ruvolo, Ting-fan Wu, Jacob Bergsma, and Javier Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 22, 2009. URL https://proceedings.neurips.cc/paper_files/paper/2009/file/f899139df5e1059396431415e770c6dd-Paper.pdf.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950. doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 11328–11348, 2023. doi: 10.18653/v1/2023.acl-long.634.

Wenbo Zhang, Lijinghua Zhang, Liner Xiang, and Hengrui Cai. Reasoning is not free: Robust adaptive cost-efficient routing for LLM-as-a-judge. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. arXiv:2605.10805.

Jitian Zhao, Changho Shin, Tzu-Heng Huang, Satya Sai Srinath Namburi GNVV, and Frederic Sala. CARE: Confounder-aware aggregation for reliable LLM evaluation. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. arXiv:2603.00039.

Justin Zhao, Flor Miriam Plaza-del-Arco, Benjamin Genchel, and Amanda Cercas Curry. Language model council: Democratically benchmarking foundation models on highly subjective tasks. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pp. 12395–12450, 2025. doi: 10.18653/v1/2025.naacl-long.617.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Tianxin Zhou and Ruixi Lin. JuryProbe: An empirical consensus-risk diagnostic for routing reference-free factuality judge panels to grounded verification. *Transactions on Machine Learning Research*, 2026. arXiv:2608.20607.

Bin Zhu and Yanghui Rao. A finite-calibration regime map for LLM judge panels, 2026. ARXIV:2606.01034v1.

Bin Zhu, Yi Xie, and Yanghui Rao. Stopping and routing LLM judge panels, 2026. ARXIV:2608.19802.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. JudgeLM: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations (ICLR)*, 2025. ARXIV:2310.17631.

A ROADMAP TO THE APPENDICES

The appendices that follow are organised by topic for navigability, so they do not follow the order in which the main body cites them. For ease of discoverability, we introduce them briefly here. Appendices B–E extend the account the paper is built on, from the item-level test that supports it to the transfer result that bounds it. Appendices F–H carry the comparisons against systems we did not build, and the calibrations that license them. Appendices I–M hold the per-dataset tables, the measurement protocol, and the reference material a reproduction would need.

Table 3 maps each appendix to what it carries and to the part of the body it supports, so that any headline in the main text can be traced back to the measurement behind it. No appendix is required to follow the argument; each is written to be read on its own, and the body cites it at the point its content is used.

B THE COMPETENCE/CORRELATION ACCOUNT IN FULL

This appendix carries the account of §3.3 in full: the two quantitative views of it –an item-level test with real power, and a cross-dataset predictor without it– then the regime ordering read dataset by dataset, the member-level views, the panel-size sweep, and the effective-independence comparison the body only points at.

Item-level recovery test (confirmatory). The mechanism predicts that the panel beats the best single by recovering the leader's errors from decorrelated member competence. We test this directly: pooling the eight datasets and restricting to the items on which the CV-best single judge is wrong –so that any recovery must come from the other members– we relate the panel's per-item recovery to the fraction of members individually correct on that item (Table 4). The relation is strong and orderly (Table 4(a)): the panel never recovers an item on which every member shares the leader's error (0 of 529 such items), recovers about one in six once two members dissent, and about six in seven where nine of the ten are right. A logistic regression with dataset fixed effects gives a coefficient of $+4.67$ (log-odds) on the members-correct fraction, and the correlation between recovery and that fraction –a Pearson correlation between the binary recovery outcome and the continuous fraction, hence a point-biserial one– is $+0.41$ pooled over the $1{,}998$ items ($p < 10^{-20}$) and positive within each of the eight datasets, median $+0.48$ (Table 4(b)). This test, and not the dataset-level correlation below, carries the mechanism's quantitative support, for two reasons. The first is that conditioning on the leader being wrong removes the leader from the test: on these items its verdict is fixed at incorrect, so a relation between recovery and the *other* members' competence cannot be an echo of how strong the leader is, nor of the selection that made it the leader. The second is that the unit of analysis is the item, so the estimate rests on nearly two thousand points instead of the eight a dataset-level correlation must rest on — and eight points cannot resolve an effect of this size, as the next paragraph shows.

Cross-dataset predictor (illustrative only). The dataset-level counterpart of that test is reported here for completeness, and is the one quantity in this appendix the paper does not lean on. The account also makes a prediction across datasets –the panel's edge should be largest where member errors decorrelate and competence is spread over several members– and it is fair to ask whether the eight datasets bear that out numerically as well as qualitatively. We form, per dataset, the product of the two quantities the account names: $1 - \bar{\phi}$, where $\bar{\phi}$ is the mean pairwise error-correlation of the ten members (the quantity Figure 4 displays in full), and the number of members reaching $\kappa \geq 0.30$, our count of how far competence is spread; these are the two coordinates of the plane in Figure 3(a).

Table 3: Roadmap to the appendices: what each one carries, and the sections of the main text it supports.

App.	What it carries	Supports
The account, its extensions, and its bounds		
B	The competence/correlation account in full: the item-level recovery test that carries it, the cross-dataset predictor that does not, the regime ordering read dataset by dataset, the member-level views, the panel-size sweep, and the effective-independence comparison against a published frontier panel	§3.3
C	The identity that makes the correlation coordinate free, the obstruction that keeps the competence one from being so, the threshold family and its check on 1,888 (sub-)panels, the labelling budget for both the test and the aggregator, and which half of the rule these eight datasets actually test	§3.3, §4
D	Matched base-versus-instruct probe: how instruction tuning moves the panel's error-correlation	§3.3
E	Leave-one-dataset-out transfer, and why the in-domain advantage does not survive it	§3.3
External comparisons and their calibration		
F	Shared rows, the exact metric conversion, and the harness calibrations behind the comparison with the published leaderboard, and the provenance of the claim-support frame as a natural experiment on member competence	§2, §3.3, §4
G	The purpose-trained detectors, run through our own harness across the whole suite rather than on their home benchmark alone	§1, §4
H	Larger open judges under the same protocol, the panel–frontier paired intervals (Table 13), and the cost model behind the quoted break-even	§4
Protocol, tables, and reference material		
I	Full per-dataset ladders on both metrics, the rescue and aggregator contrasts, and the weights the stacker learns	§2, §3
J	How each dataset's class-balanced item set and its clusters were built	§2
K	The judge roster, the prompt and decoding settings, the coverage accounting, and the multiplicity audit	§2, §3
L	Extended related work: judges, panels of them, cascades, purpose-trained specialists, and the benchmarks	§1, §3, §4
M	The scope conditions beyond the roster limitation the conclusion names	§4, §5

Rank-correlating that composite against each dataset's (stacked $-$ best-single) κ gap gives Spearman $\rho = +0.47$ — the predicted sign, and consistent with the regime ordering read off in §3.3, but not significant ($p = 0.24$). Nor could it easily be, since at $n = 8$ only a strong ordering ($\rho \gtrsim 0.74$) would register, and a composite that squeezes two quantities into one number cannot be expected to produce one.

Table 5 reads the two coordinates separately, which is the more informative view and the less flattering one. The competence coordinate carries the stronger association in every subset ($+0.48$ to $+0.74$), and the correlation coordinate the weaker one (-0.36 to -0.61). Both, however, are sensitive to the obvious check of dropping the sets held back in §2, yet most of that sensitivity comes from one dataset. CNN, in fact, carries the most correlated panel in the suite ($\bar{\phi} = 0.854$, or an effective $n_{\text{eff}} = 1.15$ independent votes on the scale defined below) together with the third-largest gap ($+0.051$, behind only FACTSCORE and WICE), which is the opposite of what the account predicts. It is also the set held back for having too few clusters to resolve any paired difference, so that gap is a central estimate whose interval accommodates either sign. WICE carried a distortion of another kind, one that re-framing the four LLM-AGGREFACT sets removed. Under the misframed prompt none of its members was competent, yet the stacker beat the best single judge by $+0.116$, by correcting a

Table 4: Item-level mechanism test on the 1,998 items on which the CV-best single judge is wrong. Items are pooled over the eight datasets, so that any recovery must come from the *other* members. **(a)** Recovery probability against the number of members individually correct on the item, counting only those that returned a verdict; **(b)** the same relation within each dataset.

members correct	P(recovers)	n
0	0.000	529
1	0.058	327
2	0.162	235
3	0.101	189
4	0.157	210
5	0.225	160
6	0.310	145
7	0.374	91
8	0.468	77
9	0.857	35

(a) by members correct

Dataset	n	P(recovers)	r
MEDHALLU	253	0.316	+0.75
RAGTRUTH	410	0.173	+0.16
TRUTHFULQA	431	0.074	+0.30
FACTSCORE	89	0.303	+0.62
WICE	57	0.228	+0.62
XSUM	163	0.153	+0.53
CNN	40	0.100	+0.43
EXPERTQA	555	0.068	+0.37

(b) by dataset

r: correlation between recovery and the fraction of members correct on the item.

Table 5: Cross-dataset predictor of the panel's gain over its best single judge. Each coordinate is taken separately and as the composite; rows drop, in turn, the two sets held back in §2.

		Spearman ρ (two-sided p) against the (stacked − best-single) κ gap		
subset	n	$\bar{\phi}$	members $\kappa \geq 0.30$	$(1-\bar{\phi})\times$ members
all eight	8	−0.36 (0.39)	+0.48 (0.23)	+0.47 (0.24)
without CNN	7	−0.61 (0.15)	+0.74 (0.06)	+0.71 (0.07)
without CNN and ExpertQA	6	−0.37 (0.47)	+0.58 (0.23)	+0.54 (0.27)

n: number of datasets left in the subset.

miscalibrated leader rather than by pooling complementary competence (Appendix F); under the corrected frame, nine of its ten members clear $\kappa = 0.30$. The re-framing thus removed a confound, though not the lack of power.

No entry in Table 5 reaches significance at $n \leq 8$. Read instead against the panel's *absolute* stacked κ, rather than against its gap over the best single, the two coordinates give $\rho = -0.76$ for $\bar{\phi}$ and $\rho = +0.95$ for the member count; both, however, are confounded by how hard the task is, and neither is a statement about the panel's advantage. The competence/correlation account does not depend on this reading, its quantitative test being the item-level one of Table 4, over some two thousand points rather than eight. We therefore read the dataset-level axis qualitatively, through the regime ordering and the plane of Figure 3(a).

The regime ordering by dataset. The regime ordering that §3.3 reads off Figure 3(a) places the eight datasets in four groups. The improvement over the best single is largest where competence is distributed over several capable members whose errors decorrelate (FACTSCORE, MEDHALLU, WICE); real but unresolved where the members carry moderate, mutually redundant competence (XSUM); negligible where one model dominates a weak field and cross-validation reliably finds it (RAGTRUTH and TRUTHFULQA, in each of which stacking simply recovers `Gemma`); and unresolvable either way where competence is scant and the errors are the most correlated in the suite (CNN, EXPERTQA). That ordering carries two qualifications. The first concerns power rather than effect. The point gain over the best single is largest on the three least-powered sets, FACTSCORE (+0.059 at $n = 330$), WICE (+0.056 at $n = 222$), and CNN (+0.051 at $n = 114$), all above MEDHALLU's clear +0.032; their paired intervals, however, run two to three times wider and cannot exclude zero, so the one resolved win bounds the panel's advantage from below, and the favourable end of the ordering is under-resolved rather than inactive. The second concerns the contrast the literature would predict. Grounded-versus-ungrounded is numerically present in the suite yet mechanistically empty: FACTSCORE is the most panel-favourable dataset we measure, whereas TRUTHFULQA is a textbook dominant-leader set whose difference against the best single is indistinguishable from the grounded RAGTRUTH's. The two ungrounded sets thus occupy opposite regimes, and "ungrounded" is not a coherent behavioural category for this question.

Member-level views. Two member-level figures complete the account, and both are to be read by regime, not member by member.

Figure 4 gives the pairwise error-correlation structure of the panel on all eight datasets, ordered by $\bar{\phi}$, and the progression across the eight is the axis the account rests on. On FACTSCORE ($\bar{\phi} = 0.11$) the members are close to independent and several pairs are mildly *anti*-correlated, so nearly every member carries information the others lack; by RAGTRUTH (0.55) whole blocks of the matrix have gone dark and only a few pairs remain informative; and on CNN (0.85) the panel is close to one judge repeated ten times, which is why aggregation has so little to exploit there whatever the members' individual competence. The figure also shows the two axes of §3.3 to be distinct: TRUTHFULQA sits at the decorrelated end (0.26) and is nonetheless a dominant-leader set whose difference against the best single does not resolve, because only two of its members are individually competent, so low correlation buys nothing without competence to decorrelate.

Figure 5 reads the same regimes member-wise, by refitting the stacker without each member in turn. Where the panel's improvement is largest, the profile is flat: on MEDHALLU no member is worth more than 0.006 κ to the panel, and on FACTSCORE five members each cost 0.03 or more, so the gain is jointly held rather than owed to any one judge. Where a single judge dominates, one cell carries the column: dropping `Gemma` costs 0.066 on TRUTHFULQA and 0.063 on RAGTRUTH, against at most 0.018 for any other member; on WICE it costs 0.062 against 0.036 for the next member, a leader standing out of a competent field rather than dominating a weak one. The extreme case is CNN, where dropping `Gemma` costs 0.158 — by some margin the largest effect in the figure, and a reminder that a 114-item set with the suite's highest error-correlation is carried by whichever member happens to be competent on it. No removal helps by more than $+0.03$, so a blue cell marks a member the panel can spare rather than one it would be better without. Read as a rule, a flat profile signals distributed competence and predicts the larger improvement over the best single, whereas a single deep notch signals a dominant leader and predicts an unresolved one; in neither case is the gain the contribution of one strong member alone.

Panel size. Sweeping the panel from one to ten members corroborates the same ordering directly (Figure 6(a), members added in a greedy decorrelating order and the stacker refitted at each K). Competence keeps accruing with each added member on the low-correlation sets (MEDHALLU, FACTSCORE); the leader-dominated (RAGTRUTH, TRUTHFULQA) and the redundant (XSUM) sets stay flat; and WICE peaks early before diluting. Aggregation therefore rewards decorrelated competence and not member count, so we read the sweep as a diagnostic and not as a recipe for choosing K.

Effective independence against a frontier panel. The correlation coordinate has a standard summary, which lets us place our panel against one we did not run (§3.3). Writing $\bar{\phi}$ for the mean pairwise error-correlation over the $\binom{k}{2}$ member pairs (the correlation coefficient amongst the pairwise diversity measures of Kuncheva & Whitaker (2003), computed on correct/incorrect outputs and hence identical on error indicators), Kish's effective sample size $n_{\text{eff}} = k/(1 + (k-1)\,\bar{\phi})$ gives the number of independent judges a panel of k correlated ones is worth. Our ten members are worth $n_{\text{eff}} = 5.11$ independent votes on FACTSCORE, 2.98 on TRUTHFULQA, 2.58 on WICE, 2.49 on MEDHALLU, 1.84 on XSUM, 1.69 on RAGTRUTH, 1.64 on EXPERTQA, and 1.15 on CNN, which places the correlation axis of the account on a single scale. Kohli (2026) evaluates the same $\bar{\phi}$ statistic for a panel of *nine frontier judges* on three natural-language-inference (NLI) corpora, reporting n_{eff} between 2.18 and 2.48, with the best single judge unbeaten and an asymptote at $1/\bar{\phi} \approx 2.6$ read there as a bound on what panels of current models can reach. No row is shared between the two studies and the tasks differ, so this compares a derived, panel-internal quantity rather than a score; yet on that quantity our cheap panel exceeds their frontier panel's best on four of eight datasets and passes the asymptote on two, reaching nearly twice that value on FACTSCORE.

In concurrent work on preference data, Hossain et al. (2026) claim the same ordering on a shrinkage estimate of $\bar{\phi}$, ten open judges being worth 3.5 independent ones and three frontier judges 1.4. We interpret this as the prediction of Kim et al. (2025) and Goel et al. (2025) realised –error-correlation rises with model accuracy, so a panel assembled from weaker and more heterogeneous judges has more independence to aggregate– and recover Kohli's negative result: a panel of nine strong, mutually redundant judges sits just above our four least favourable sets on that scale, and on those four we too

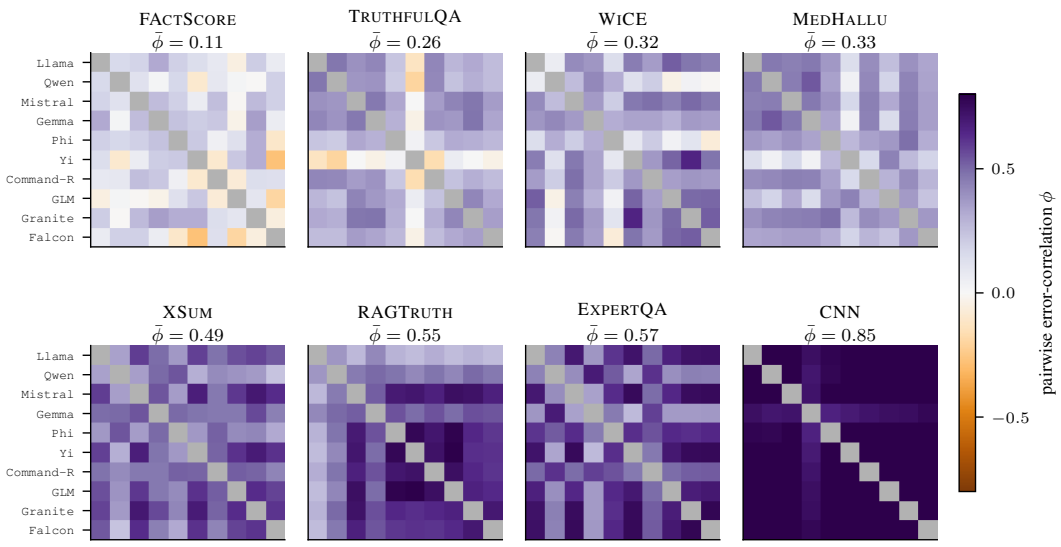


Figure 4: Pairwise error-correlation ϕ between the ten members, on all eight datasets in ascending $\bar{\phi}$ (panels share the member order and the colour scale; the diagonal is masked). The panels run from FACTSCORE (lowest $\bar{\phi}$) to CNN (highest).

find aggregation recovers the leader and no more. It also shows the asymptote to be a property of that roster rather than of panels, since it is $\bar{\phi}$ that fixes the bound and $\bar{\phi}$ is a choice of members.

Finally, Li & Hai (2026) argue the same dependence analytically for committees of language models: under a one-factor model, positively correlated errors leave every rule that reads only the votes a positive error floor however large the committee, and reduce a committee of fixed size to single-member behaviour as the correlation approaches one. The two axes remain distinct here as elsewhere: TRUTHFULQA carries $n_{\text{eff}} = 2.98$ and still yields nothing, only two of its members being individually competent.

C THE ADMISSIBILITY TEST AND ITS LABELLING BUDGET

This appendix carries the measurement behind §3.3: the rule itself, the identity that makes one regime coordinate free, the obstruction that keeps the other from being so, and the budget at which the call stabilises. Figure 6 collects the two budgets a deployer sets by dataset, the panel size of Appendix B and the calibration size measured here.

The rule. The test reads two statistics off a labelled calibration sample, both computed from the members' own verdicts and neither from the stacker's weights. The first is the competence spread $s = |\{i \mid \kappa_i \geq \kappa_0\}|$, the number of members whose own κ against gold clears a competence bar κ_0; the second is $\bar{\phi}$, the mean over the $\binom{N_v}{2}$ member pairs of the Pearson correlation between their error indicators $e_i = \mathbf{1}[v_i \neq y]$, each pair read on the items both members committed on. For a competence bar κ_0, a share c, and a correlation bar ϕ_0, a task is *admissible* when $s \geq c\,N_v$ and $\bar{\phi} \leq \phi_0$. The numbers we report are those of one instance, $(\kappa_0, c, \phi_0) = (0.30, \frac{1}{2}, 0.40)$, declared once and applied unchanged to every dataset and every calibration sample, the competence bar being the one of §3.3 and Figure 3. These values were fixed with the eight results in view, yet not tuned: eight datasets are too few for any tuning to be stable, as the next paragraph shows, and the paper therefore relies on what holds across the family rather than on the instance itself.

In use, admissibility decides whether to fit the stacker rather than select a single member, and on this suite every admitted task is also one on which the frontier judge is not materially ahead; whether the frontier judge is worth keeping on a task that fails it is read, qualitatively, from the same two coordinates, as the dominant-leader corner of the plane (§4). On the full samples the test admits MEDHALLU ($s = 10$, $\bar{\phi} = 0.335$), WICE (9, 0.320), and FACTSCORE (5, 0.106), and rejects the

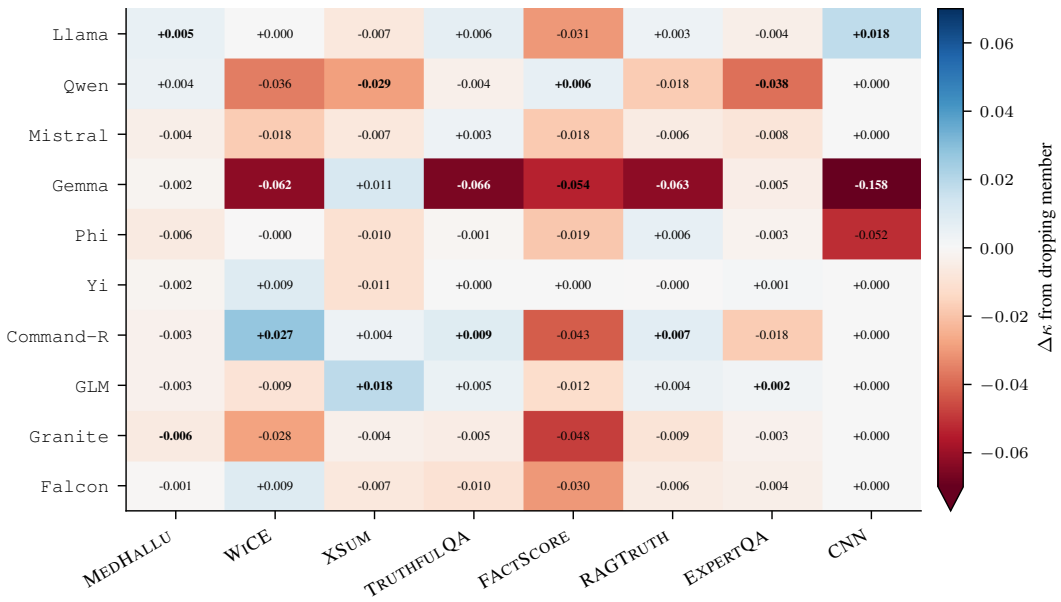


Figure 5: Leave-one-out change in stacked κ from dropping each member (Δ = stacked without member − full stacked): red marks a load-bearing member, whose removal costs the panel, blue one the panel can spare. Every cell carries its value, and each column's extremes are set in bold. The colour scale is clipped at ± 0.07 for legibility, `Gemma` on CNN at -0.158 being off-scale.

other five (Table 6); the suite's one clear improvement over the CV-best single, on MEDHALLU, falls inside the admissible set, and no rejected task carries one.

It is however worth stating that the test itself is a decision heuristic rather than a hypothesis test, and carries no interval of its own. We measure its reliability below instead, as the probability that a calibration sample of a given size reproduces the full-sample call, and name there its two failure cases, the boundary sets FACTSCORE and TRUTHFULQA.

The rule as a family of thresholds. Two properties of the rule, and not of our data, make its thresholds a choice rather than an estimate. First, the admissible set is monotone in each threshold: raising κ_0 or c, or lowering ϕ_0, can only remove tasks from it. The family therefore runs from a screen that admits every task to one that admits none. As such, choosing a member of it means choosing an operating point between two errors, i.e. dropping the frontier judge where it was needed or paying for it where it was not. Their relative cost, consequently, belongs to the deployer, and the data cannot estimate it.

Second, on a finite suite the calls are piecewise constant in the thresholds, changing only where a threshold crosses some dataset's own coordinate. Any criterion scored on the suite is therefore optimised by a cell of that partition rather than by a point, and which cell wins is decided by the datasets on its edges. At $\kappa_0 = 0.30$, our instance lies in a cell of the plane spanned by the required count $\times$ ϕ_0 that is L-shaped rather than rectangular (hatched area in Figure 3(a)). The count may run from three to five members and ϕ_0 from 0.335 to 0.494, the upper limit becoming unbounded at a count of five, since no rejected dataset has five competent members. Its edges are set by the datasets they touch: MEDHALLU's $\bar{\phi}$ on the left, TRUTHFULQA's two competent members below, FACTSCORE's five above, and XSUM at the inner corner. The competence bar moves the points of that plane rather than an edge, and with the other two thresholds fixed, the result is unchanged for $0.269 \leq \kappa_0 \leq 0.339$, between the fifth-ranked members of TRUTHFULQA and FACTSCORE; no call in the paper changes within either range.

The sub-panels of our ten members show both properties at work. Each of the $1{,}888$ (sub-)panels we draw on the same items (per dataset, every subset of eight or nine members, sixty drawn of five, six, and seven, and the full ten) has its own two coordinates, which we score with the paper's stacker against its own CV-best member and against the frontier judge. We score a threshold pair by Youden's

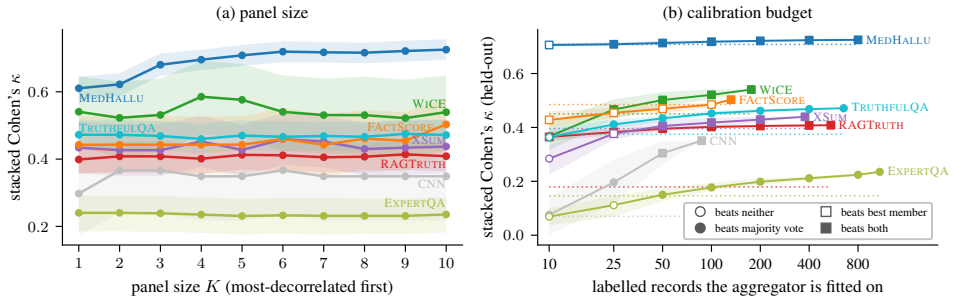


Figure 6: The two budgets a deployer sets by dataset. **(a)** Stacked κ against panel size K, members added in a greedy decorrelating order and the stacker refitted at each K (Appendix B). **(b)** Stacked κ against the labelled *records* the aggregator is fitted on, drawn within the reported stacker's own folds and pooled across them, interquartile range over $R = 100$ draws shaded. The dotted line reports the dataset's label-free majority vote on the same folds; markers encode what a budget beats.

index (Youden, 1950), $J = \mathrm{TPR} + \mathrm{TNR} - 1$, the rate at which the test admits the panels that meet a target plus the rate at which it rejects those that do not, less one (equivalently, $2\,\mathrm{BACC} - 1$ of the call against the target), so that $J = 0$ for a call unrelated to the target and $J = 1$ for a perfect one. Selecting c and ϕ_0 on the panels by J, with the target that the panel stays within 0.10 κ of the frontier judge, gives $c = 0.4$ with any $\phi_0 \geq 0.53$ rather than our instance. Selected instead on seven datasets and applied to the eighth, the optimum moves from fold to fold (c from 0.3 to 0.5, ϕ_0 from 0.43 upwards), as the partition predicts, yet it returns our call on all eight held-out datasets, and on seven of eight when the target is to beat the CV-best member.

What does not move, however and most importantly, is the overall direction. We scan a grid of $c \in \{0.3, 0.4, \ldots, 0.8\}$ and $\phi_0 \in \{0.25, 0.26, \ldots, 0.90\}$, $6 \times 66 = 396$ pairs, and keep those that split the panels, admitting some and rejecting others. At every such pair $J > 0$ holds for both targets: the admitted panels are more often within reach of the frontier judge than the rejected ones, and more often above their own best member. This holds at all 396 pairs on the drawn panels, every one of which splits them, and at all 375 that split the eight full ones, the other 21 ($c \geq 0.6$ with $\phi_0 < 0.32$) admitting none of the eight; likewise, with the frontier margin tightened to 0.05 κ, it still holds at 92% of the pairs on the drawn panels, and at all of them on the full ones. We therefore offer the test as a family whose every member points the same way, and our instance as one operating point of it. The sub-panels share members and items, so this is evidence across panel configurations rather than across tasks, and a new task remains the test the suite cannot supply.

The agreement identity. We write $v_i \in \{0, 1\}$ for member i's verdict on an item carrying gold y, and $e_i = \mathbf{1}[v_i \neq y]$ for its error indicator. Since y is common to the two members being compared, $v_i = v_j$ holds exactly when $e_i = e_j$, so the pairwise agreement *rate* is the same statistic read through either lens, and averaging over the 45 member pairs preserves the equality. Member disagreement is therefore an observable proxy for error-decorrelation. The identity is not new: it is the pairwise case of the agreement equations from which Platanios et al. (2014) estimate classifiers' error rates without labels, each agreement rate being written through the two members' error rates and their joint one, whereas we use it only to make the correlation coordinate observable. The Pearson $\bar{\phi}$ of Figure 4 is the same agreement centred and scaled by the members' error rates, and it is that normalisation, not the association, which consumes gold; the unnormalised rate ranks the eight datasets as $\bar{\phi}$ does at Spearman $\rho = +0.93$ ($p = 0.001$).

The competence obstruction. The natural candidate is the Dawid–Skene fit already used as an aggregator baseline (§3.1), which estimates each member's reliability by expectation-maximisation without gold. It fails here for a structural reason. Its estimate is a member's agreement with a latent consensus, and where the panel shares an error the consensus is the error: on CNN, the suite's most correlated panel, it rates all ten members competent at $\kappa \geq 0.30$ against the consensus, whereas not one of them clears that bar against gold, and the same happens on EXPERTQA. Across the eight datasets its count of competent members rank-correlates with the gold count at $\rho = -0.01$, and it recovers the true best member on three datasets of eight — failing, in particular, on RAGTRUTH, the

suite's clearest dominant-leader set. Balasubramanian et al. (2026) claim a complementary failure on the aggregation side: on correlated judges, an aggregator that assumes conditional independence can be confidently wrong even when handed each judge's true accuracy and, under some dependence structures, it stays strictly suboptimal however large the panel grows.

Nor does a richer label-free model escape the obstruction without an assumption. The agreement rates of all subsets of the N_v members supply $2^{N_v} - N_v - 1$ equations for the $2^{N_v} - 1$ individual and joint error rates, short by exactly the N_v individual ones (Platanios et al., 2014), so every such estimator closes the system by an assumption on how the errors depend on one another –conditional independence for Dawid–Skene, the least dependence the agreements allow for Platanios et al., independence between groups of dependent members for Jaffe et al. (2016)– and by members better than chance. A panel whose errors load on one shared source and whose average member sits near chance, as on CNN ($\kappa = 0.088$), is hence precisely where those assumptions fail. An unlabelled sample can say that a panel agrees with itself; it cannot say whether the panel is right.

The budget sweep. For each dataset and each budget $n \in \{10, 25, 50, 100, 200\}$ not exceeding the dataset, we draw $R = 200$ calibration samples and recompute both coordinates and the panel's apparent leader on each. Sampling is by cluster and truncated at the cluster boundary, matching the resampling unit of §2: a deployer buys labelled records, and taking half a matched pair would give the sample a class balance the labelled set would not have. Table 6 reports how often the full-sample regime call is reproduced: on average 85.5% of the time from fifty labelled records and 91.9% from a hundred, with $\bar{\phi}$ itself recovered to within a median 0.022 at the latter. Two datasets sit at the decision boundary and are correspondingly unstable, FACTSCORE carrying the five competent members the rule requires and no more, and TRUTHFULQA's members clustering at $\kappa = 0.30$.

One property of the rule at our instance should be stated plainly. There, the correlation bar is not what rejects any dataset. Every dataset with at least five competent members also satisfies $\bar{\phi} \leq 0.40$, and every dataset above the bar has already failed the count, so at the instance the call coincides with the competence threshold alone on all eight. This is a property of eight points rather than of the rule. The bar binds from below, since a value under MEDHALLU's 0.335 would reject it, and at a share of three or four members it is the bar alone that rejects XSUM; what the suite lacks at our share is a dataset with five competent members and strongly shared errors, which is why the hatched region of Figure 3(a) runs unbounded at a count of five. We therefore report the conjunction as the account requires it, noting that the eight full panels do not test its correlation half at our instance. The sub-panels do populate that cell: 106 of the 680 with enough competent members exceed the correlation bar, on MEDHALLU, WICE, and XSUM, and on the two datasets where both sides of the bar occur the more correlated ones gain less over their best member ($+0.030$ against $+0.038$ on MEDHALLU, -0.035 against $+0.023$ on WICE), the direction the correlation half predicts, although on two datasets only. The optimum above is flat for every $\phi_0 \geq 0.53$ for the same reason: at $c = 0.4$ no drawn panel with enough competent members lies above that value.

The instability of the two boundary sets belongs to the threshold and not to the estimate: replacing the count by the continuous competence the panel carries above the bar, $\sum_i \max(0, \kappa_i - 0.30)$, and normalising each statistic by its own spread across the eight datasets, the continuous form resolves about twice as finely at a fifty-record budget (0.055 against 0.100). We keep the count in the regime plane of Figure 3(a), where it is read qualitatively and its interpretability is worth more than its efficiency, and report the continuous companion in the released artefact.

The aggregator's own budget. The sweep above prices the *diagnostic*; the aggregator has a budget of its own, and the two are measured the same way. Table 7 fits the stacker on n labelled records and scores it out-of-fold, against the best single member selected on those same records and the label-free majority vote. Three design points matter for reading it. Budgets are in labelled *records*, the clustering unit of §2, which on the paired sets is two items; a deployer buys records, not items. The folds are the reported stacker's own, and within each fold the budget is drawn from that fold's training clusters, the fit predicts that fold's test items, and the predictions are pooled before κ is taken — so the estimator is the paper's, and at the full training pool the curve reproduces the reported value to within 0.003, which is the check that the protocol has not drifted. Averaging per-fold κ, instead, would be a different and biased estimator, κ not being linear in the confusion counts, and disagrees with the pooled one at the full pool by up to 0.010, on CNN, whose training pool is the smallest.

Table 6: Recovery of the regime call from a limited labelling budget. Per dataset and budget, $R = 200$ calibration samples of that many labelled records are drawn as whole clusters (§2), both coordinates of §3.3 being re-estimated on each.

		full sample			Pr(call matches) *at a budget of*				
Dataset	n	$\bar{\phi}$	spread	regime	10	25	50	100	200
MEDHALLU	2000	0.335	10	admissible	0.70	0.78	0.88	0.96	0.99
WICE	222	0.320	9	admissible	0.54	0.67	0.81	0.97	1.00
RAGTRUTH	1362	0.546	2	not admissible	0.98	1.00	1.00	1.00	1.00
XSUM	546	0.494	4	not admissible	0.79	0.87	0.94	0.99	1.00
CNN	114	0.854	0	not admissible	0.99	1.00	1.00	1.00	—
EXPERTQA	1462	0.566	0	not admissible	0.96	0.99	1.00	1.00	1.00
FACTSCORE	330	0.106	5	admissible	0.69	0.54	0.67	0.81	—
TRUTHFULQA	1634	0.261	2	not admissible	0.36	0.51	0.55	0.62	0.76
suite: mean Pr(call)					**0.75**	**0.79**	**0.85**	**0.92**	**0.96**
suite: Pr(same leader)					0.32	0.46	0.56	0.69	0.76
suite: median $\lvert\Delta\bar{\phi}\rvert$					0.081	0.046	0.033	0.022	0.015

regime: the full-sample call, a task being admissible at a competence spread of at least five of the ten members at $\kappa \geq 0.30$ and $\bar{\phi} \leq 0.40$. Budget columns: fraction of samples reproducing that call; —: budget exceeding the dataset.

The label-free majority vote is drawn on the same folds as a dotted floor per dataset, because what a deployer needs to know is whether the labels they can afford beat spending none.

Two readings the body compresses are clearer per dataset. The budget at which the fitted stacker overtakes that vote runs from ten labelled records (RAGTRUTH, CNN) through twenty-five (MEDHALLU, WICE, TRUTHFULQA) to fifty (XSUM, EXPERTQA), and on FACTSCORE not until its full pool of 132 — the vote there being within 0.018 κ of full supervision, so there is little for labels to buy. EXPERTQA is the one set that keeps paying throughout, rising from $\kappa = 0.069$ at ten records to 0.235 at its full pool of $1{,}088$. That ordering is not the regime ordering of §3.3: the sets on which a handful of labels already pays are the rejected RAGTRUTH and CNN, and the admissible FACTSCORE is the slowest, its free vote being already within reach of full supervision. The two alternatives are not nested, and neither dominates the other. The free vote is the harder bar on MEDHALLU and FACTSCORE at every budget, on WICE below fifty records and on XSUM below a hundred; everywhere else, including on WICE and XSUM once their thresholds are passed, the harder bar is the best member selected on the same labels. This is why Figure 6(b) encodes the two on separate channels rather than on one ordinal scale.

The two headline budgets quoted in the body summarise that table, and each has its own exceptions. Fifty labelled records suffice to clear the label-free vote on seven of the eight datasets, FACTSCORE excepted for the reason above; and a hundred bring the stacker within 5% of its fully supervised κ on six of the eight, the exceptions being EXPERTQA, which is still 25% short at that budget and keeps paying to its full pool, and CNN, whose training pool is too small for the hundred-record point to exist. Below the crossing point a calibration set is worse than none: at ten labelled records the fitted stacker sits *beneath* the free vote on six of the eight sets, all but RAGTRUTH and CNN, so a deployer who cannot reach the crossing is better served by not fitting an aggregator at all.

D BASE-VERSUS-INSTRUCT PROBE

This appendix gives the procedure behind the finding of §3.3 on where the panel's error-correlation comes from. To separate a shared answer-bias from shared task-difficulty as the source of the panel's error-correlation, we re-run the eight panel families that release a public base (pre-instruction-tuning) checkpoint, i.e. all members except `Phi` and `Command-R`, which ship only in aligned form. Each is run in both its base and its instruction-tuned variant under a matched constrained-decoding scorer, holding the underlying models fixed. For each variant we compute the within-panel mean pairwise error-correlation (ϕ, Pearson on per-item error indicators) across all eight datasets; the two distributions are plotted in Figure 7. The shift $\Delta\phi$ (instruct$-$base) is negative on seven of the eight datasets, ranging from -0.085 on WICE to -0.495 on FACTSCORE, and null on MEDHALLU ($+0.018$, interval including zero); the base panel's ϕ reaches 0.667 on RAGTRUTH. Excluding

Table 7: Gold-label learning curve of the aggregator: mean Cohen's κ over $R = 100$ draws per budget, budgets being in labelled *records*, the clustering unit of §2. Folds are the five of the reported stacker, the budget drawn from each fold's training clusters and the out-of-fold predictions pooled before κ is taken.

Dataset	train	unw.	*labelled records the stacker is fitted on*							
			10	25	50	100	200	400	800	full
MEDHALLU	800	0.708	0.706	0.709	0.713	0.718	0.721	0.724	0.725	0.725
WICE	176	0.450	0.367	0.466	0.502	0.521	—	—	—	0.541
RAGTRUTH	544	0.179	0.365	0.383	0.395	0.402	0.405	0.408	—	0.408
XSUM	377	0.396	0.284	0.377	0.405	0.418	0.429	—	—	0.440
CNN	87	0.070	0.076	0.196	0.304	—	—	—	—	0.351
EXPERTQA	1088	0.146	0.069	0.112	0.150	0.177	0.199	0.211	0.224	0.235
FACTSCORE	132	0.485	0.428	0.454	0.469	0.485	—	—	—	0.503
TRUTHFULQA	653	0.375	0.364	0.411	0.434	0.452	0.462	0.468	—	0.471

train: smallest per-fold training pool, capping the grid; *unw.*: label-free majority vote on the same folds; *full*: the whole pool, reproducing the reported cross-fitted value to within 0.003; —: budget beyond that pool.

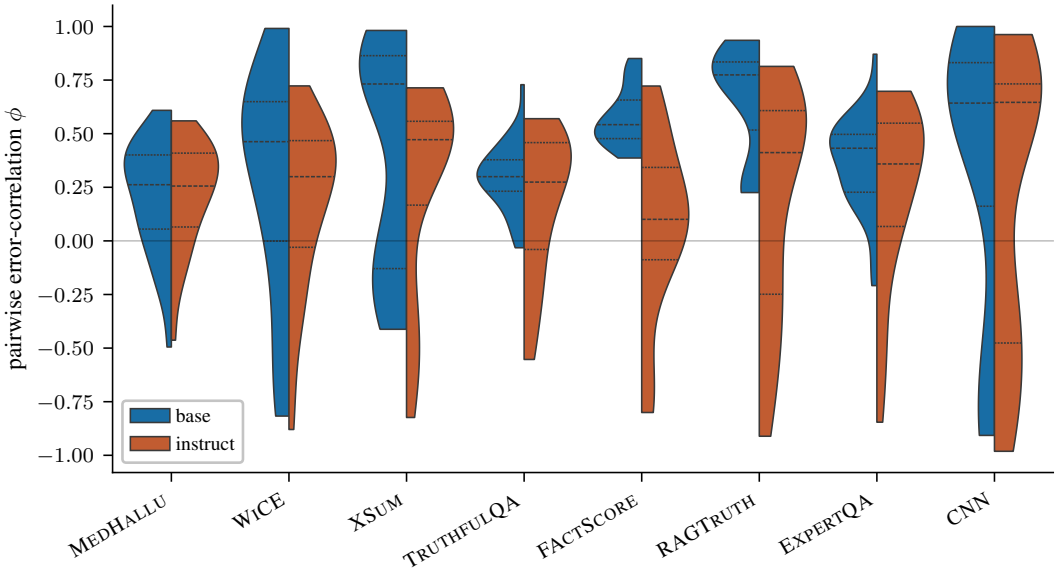


Figure 7: Within-panel pairwise error-correlation ϕ (the 28 member pairs) for the matched base and instruct panels, per dataset (each dataset under its own scoring frame, §2; constrained decoding). The shift $\Delta\phi$ (instruct−base) is summarised in Appendix D.

from both panels the one instruct member that collapses to a single class under the constrained scorer (`Falcon`), the shift remains clearly negative on six of the eight datasets, XSUM's becoming unresolved ($+0.027$ $[-0.021, +0.073]$) and MEDHALLU's small positive shift resolving ($+0.046$ $[+0.023, +0.067]$). Instruction tuning thus lowers the panel's error-correlation rather than raising it: the correlation the aligned panel carries is already present, and larger, in the base checkpoints, which is the sense in which §3.3 reads it as inherited from pretraining rather than induced by alignment. Being computed on eight members under constrained decoding, the probe's ϕ is not the ten-member $\bar{\phi}$ of Table 6, and only its shift is read here.

E TRANSFER TO AN UNSEEN DATASET

The in-domain gain raises the deployment question directly: can a panel calibrated on some datasets be deployed on another without fresh labels? It cannot, unchanged, and this appendix gives the measurement, the mechanism, and the boundary of the negative. The conclusion we draw from it

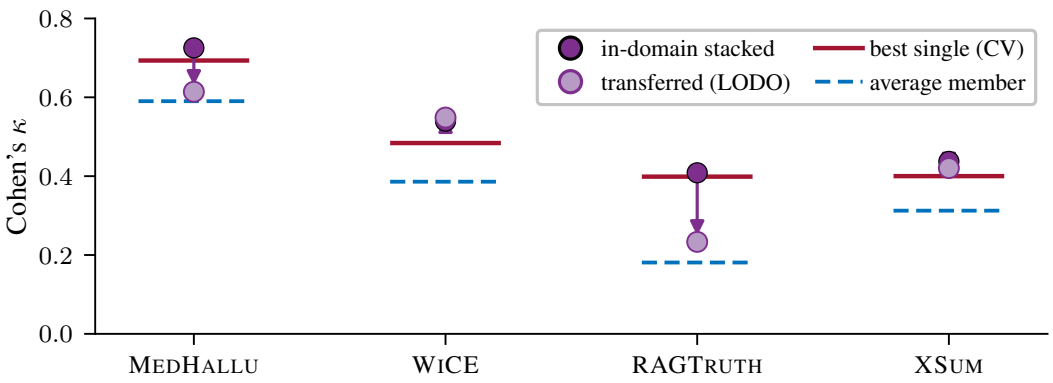


Figure 8: Leave-one-dataset-out transfer on the four core grounded sets: in-domain stacked and transferred (LODO), against the CV-best single judge and the average member per dataset.

is a scope condition on §4's costing –the calibration budget is per domain– rather than a result we develop further.

Measurement and mechanism. We test with a pooled stacker under a leave-one-dataset-out (LODO) protocol on the four core grounded sets, fitting on the other three and evaluating on the held-out set (Figure 8). Off-domain, the panel returns to roughly the level of the label-free baselines and wins nothing: zero of four transfer wins, with a significant transfer cost –the paired (in-domain − transferred) gap– on the two question-answering sets, MEDHALLU ($+0.111$) and RAGTRUTH ($+0.175$). MEDHALLU's clean win is lost –the transferred stacker drops below even the best single (-0.079)– and RAGTRUTH's unresolved in-domain improvement becomes an outright transfer loss (-0.165). On WICE and XSUM the transferred stacker shows no clear transfer cost and does not fall below the best single, though the sample cannot rule out a cost there either.

The mechanism is the one that explains the in-domain pattern: the pooled stacker learns to trust the members that are reliable *on average* over the training pool, but the most reliable judge changes by dataset, so the fixed off-domain weight vector systematically under-weights the held-out set's specific leader. On MEDHALLU the in-domain leader is `Granite`, yet the off-domain fit gives `Granite` little weight against `Gemma`, so the transferred stacker lands even below MEDHALLU's own best-on-average single. The apparent MEDHALLU balanced-accuracy gain on transfer is a coverage artefact: it beats only the abstaining CV-best baseline, and loses to the non-abstaining label-free baselines. The negative therefore reinforces the mechanistic account of §3.3: the panel helps by dataset-specific competence re-weighting, and a fixed weight vector is what that mechanism predicts should fail.

Scope of the negative result. The measurement above is taken in one setting, and the scope matters for what may be concluded from it. That setting is the generative panel, read through hard verdicts, with a stacker pooled over the other three datasets. It says that a weight vector fitted on other domains does not carry the in-domain gain, and it identifies why — the leader moves, and a fixed vector cannot follow it. It does not say that no panel transfers. In particular, the members are read here through the interface a closed judge would also expose, a verdict rather than a graded confidence, so whether a panel whose members are put on a common scale across domains transfers better is a question this setting cannot reach and we do not take up. We accordingly take the conservative reading, that the calibration budget of §4 is per domain, and leave a calibrated-panel transfer study to future work.

F COMPARABILITY WITH PUBLISHED RESULTS

The comparison of §4 against the published results needs three things to be legitimate: a shared set of rows, a shared metric, and evidence that our harness and the leaderboard's measure the same thing on them. This appendix supplies all three, states what the comparison cannot reach, and reports the two sets whose published anchors exist but sit on a different denominator.

Table 8: The four LLM-AGGREFACT sets in leaderboard shape (balanced accuracy), the only sets on which we score the same rows as the published results. Upper block published; lower block measured here under the one protocol and per-dataset frame of §2.

System	size	CNN	XSUM	WICE	EXPERTQA	mean	source
published, from the leaderboard and the cited papers							
`Bespoke-MiniCheck-7B`	7B	65.5	77.8	83.0	59.2	71.4	leaderboard
`MiniCheck-Flan-T5-L`	0.8B	69.9	74.3	72.2	59.0	68.8	Tang et al., 2024
`FactCG-DeBERTa-L`	0.4B	70.1	73.9	74.2	59.1	69.3	Lei et al., 2025, leaderboard
`HalluGuard-4B` †	4B	70.7	75.5	80.5	59.4	71.5	Bergeron et al., 2026
`Granite Guardian 3.3`	8B	67.0	74.9	76.6	59.6	69.5	leaderboard
`GPT-4o`	—	68.1	76.8	78.5	59.6	70.8	Lei et al., 2025, leaderboard
`Claude-3.5 Sonnet`	—	67.6	75.1	77.7	60.9	70.3	leaderboard
`Qwen2.5-72B-Instruct`	72B	63.6	73.0	80.2	60.1	69.2	leaderboard
`SummaC-ZS`	0.4B	51.1	61.5	62.8	55.2	57.6	Laban et al., 2022, scored by Tang et al., 2024
this work, one protocol and one frame per dataset across all eight sets							
average panel member	4–9B	54.2	65.0	63.9	55.2	59.6	
best single member (CV-selected)	4–9B	64.9	69.9	73.9	62.0	67.7	
`MiniCheck-Flan-T5-L` *(our run)*	0.8B	67.5	74.2	73.0	58.7	68.4	
`Bespoke-MiniCheck-7B` *(our run)* §	7B	55.2	63.4	60.8	53.9	58.3	
`Qwen2.5-72B` AWQ *(our run)*	72B	67.5	72.3	81.5	60.1	70.4	
stacked cheap panel (ours)	10×4–9B	67.5	71.9	77.0	61.8	69.5	
SONNET *(our run)*	—	73.5	70.9	77.4	58.2	70.0	

source: where each published value is found, *leaderboard* being the benchmark's public leaderboard (`https://llm-aggrefact.github.io/`, retrieved 2026-07-08), which carries systems post-dating the papers, and a citation being given only where the cited work itself reports the four values; †: run by its own authors and reported in their paper, not on the leaderboard; §: a checkpoint we do not reproduce (58.3 against 71.4 published), whose published row we read instead (Appendix F).

Shared rows and metric conversion. Four of our grounded sets (CNN, XSUM, WICE, EXPERTQA) are drawn from the test split of the LLM-AGGREFACT benchmark (Tang et al., 2024), whose public leaderboard[5] reports balanced accuracy per dataset, so published methods and ours are scored on the same underlying rows. On a class-balanced set, the chance-agreement term of Cohen's κ is exactly 0.5 whatever the judge's marginal, so for a full-coverage judge $\kappa = 2\,\text{BACC} - 1$ holds as an identity, not an approximation; abstention perturbs it only through the committed-set treatment of Appendix K. Our class balancing subsamples the majority class, which leaves the expected per-class recall –hence balanced accuracy– unchanged, so a published BACC is comparable to ours in expectation; F1 and precision are prevalence-bound and are, conversely, not.

Harness calibration. A shared metric on shared rows still leaves the two harnesses free to disagree, so Table 8 carries two systems run on both sides. Re-running `MiniCheck-Flan-T5-L` through the official MiniCheck package on our instances gives 68.4 mean balanced accuracy against its published 68.8, and re-running `Qwen2.5-72B` through our own prompting harness gives 70.4 against a published 69.2 for the unquantised model. Both agree to about a point, in opposite directions, which is the resolution at which we read the table: differences under three points between blocks are not informative. A second and independent calibration comes from MEDHALLU, where three of the published detectors are members of our own panel (§4): our runs of them agree with their published hallucinated-class F1 to within 2.6 points, on a different benchmark and in a different metric from the two rows above. The exception is `Bespoke-MiniCheck-7B`, which we do not reproduce (58.3 against a published 71.4) and whose published row we therefore read instead of ours: its package scores that checkpoint through a generated-token probability rather than through the classifier head its `Flan-T5` sibling uses, and the loss shows as a drop in area under the receiver operating characteristic curve (AUC-ROC) on the very same items, so we attribute it to that scoring path rather than to our protocol and do not pursue it here.

The other four benchmarks. Published work reports on the remaining four benchmarks too, and we record where the field stands on them: on MEDHALLU Table 9 matches the metric and compares levels, and on RAGTRUTH it matches the metric but not the prevalence, whereas on TRUTHFULQA and FACTSCORE no published number is a baseline we could be scored against. Tables 9 and 10 give the strongest anchor we found per set, the metric its authors report, and the structural difference that separates it from ours. However, the difference changes from set to set, and is never a matter of prompting alone.

[5] `https://llm-aggrefact.github.io/`, retrieved 2026-07-08.

Table 9: Hallucinated-class F1 on MEDHALLU and RAGTRUTH, published and measured here. Published rows are transcribed from the cited papers (RAGTRUTH: response-level F1 on its QA subtask, which we adopt); ours are computed on our balanced construction (Appendix J), an abstention counting as a missed detection. Slices differ: MEDHALLU's published table pools the 10,000-pair superset against our 1,000-row `pqa_labeled` slice, and RAGTRUTH's sits on its QA test split, at a hallucinated rate of 0.178 against our 0.5, so its two blocks are not comparable in level, F1 moving with prevalence. Three published MEDHALLU detectors are members of our panel (§4).

System		F1
MEDHALLU — *published*		
`GPT-4o` (Pandit et al., 2025)	frontier, prompted	87.7
`Qwen2.5-14B` (Pandit et al., 2025)	single open judge	85.2
`GPT-4o-mini` (Pandit et al., 2025)	budget API	84.1
`Qwen2.5-7B` (Pandit et al., 2025)	*our panel member*	83.9
`Gemma-2-9B` (Pandit et al., 2025)	*our panel member*	83.8
`Llama-3.1-8B` (Pandit et al., 2025)	*our panel member*	79.7
measured here, no member trained on the corpus		
best member by F1 (`Gemma`)	ours	84.3
stacked cheap panel	ours	**86.2**
SONNET	ours	89.0
RAGTRUTH — *published*		
`RAG-HAT` (`Llama-3-8B`) (Song et al., 2024)	trained on the RAGTRUTH train split	74.8
`LettuceDetect-large` (Kovács & Recski, 2025)	trained on the RAGTRUTH train split	70.2
fine-tuned `Llama-2-13B` (Niu et al., 2024)	trained on the RAGTRUTH train split	68.2
`GPT-4-turbo`, prompted (Niu et al., 2024)	prompt-only frontier	45.6
SelfCheckGPT (`GPT-3.5`) (Niu et al., 2024)	prompt-only, resampling	43.7
measured here, no member trained on the corpus		
best member by F1 (`Llama`)	ours	62.5
stacked cheap panel	ours	**66.3**
SONNET	ours	76.2

On MEDHALLU the published detection table pools the full 10,000-pair superset and reports hallucinated-class F1, whereas we use the 1,000-row `pqa_labeled` slice (Pandit et al., 2025); the levels nonetheless line up, `GPT-4o`'s 87.7 with knowledge against our SONNET's 89.0 and our panel's 86.2 in the same metric, and three of the detectors reported individually there (`Qwen2.5-7B` at 83.9, `Gemma-2-9B` at 83.8, `Llama-3.1-8B` at 79.7) are members of our own panel, so a reader can compare the two tables member by member. On RAGTRUTH every strong published detector is trained on that corpus's own training split and scored on its original test split at its own prevalence, whereas ours is a paired, balanced construction from the processed corpus's training split (Niu et al., 2024; Song et al., 2024; Kovács & Recski, 2025). That test split runs at a hallucinated rate of 0.178 against our 0.5, and F1 moves with prevalence, so Table 9 sets the two side by side without ranking them. Read on the question-answering (QA) subtask we take, the prompt-only anchors there are far weaker (`GPT-4-turbo` at F1 45.6) than the trained ones (`RAG-HAT` at 74.8), which is the in-domain-supervision axis of this paper appearing in someone else's numbers. On TRUTHFULQA the canonical automated judge is fine-tuned on in-domain human ratings of *model generations* and reaches 90–96% validation accuracy on that task, whereas we score the dataset's own reference answers zero-shot (Lin et al., 2022). Finally, on FACTSCORE the published estimators are retrieval-augmented by construction, approximating the human score to within 2% at corpus level, whereas we supply no evidence by design (Min et al., 2023).

Read together, the four make one point, and it is the paper's own. Wherever a published system clearly leads us, it has either been trained on the target corpus or been given evidence we withhold, and the prompted frontier anchor on MEDHALLU lands where our frontier judge lands.

A last caveat covers all eight sets: every published number arrives under its own prompting and inference protocol, and factuality-metric comparisons are known to be fragile to exactly such differences (Godbole & Jia, 2025). Our absolute levels embed one frame per dataset (`def_on`, or the claim-support frame on the LLM-AGGREFACT sets), applied uniformly to all ten members and the frontier judge (§2), which the within-suite contrasts control for but cross-paper comparisons do not.

Table 10: Published anchors on the two benchmarks without a comparable published baseline: per set, the strongest anchor we found, the metric its authors report, and the structural difference that blocks the comparison. MEDHALLU and RAGTRUTH, reported in hallucinated-class F1, are in Table 9.

	published anchor				*ours* (BACC)	
Dataset	strongest reported system	value	metric	what blocks a direct comparison	panel	SONNET
TRUTHFULQA	`GPT-judge` (fine-tuned `GPT-3-6.7B`) Lin et al., 2022	90–96	accuracy	judges model *generations* after fine-tuning on in-domain human ratings; we score the dataset's own reference answers zero-shot	73.6	81.1
FACTSCORE	retrieval-augmented estimator Min et al., 2023	—	system-level error	retrieves Wikipedia evidence by construction, approximating the human score to within 2% at corpus level; we deliberately supply no evidence and score per claim	75.2	77.6

ours: balanced accuracy on our construction (Appendix J), *not* commensurable with the anchor beside it.

Summary of the comparison. Two readings survive the caveats, beyond the placement recorded in §4. EXPERTQA's low-headroom status is corroborated externally: the leaderboard's best sits at a BACC of 60.9 across everything up to `Claude-3.5` and `Llama-405B`, and our panel reaches that ceiling (61.8) without beating its own best single, so the low-headroom verdict of §2 is a property of the dataset rather than of our panel. An independent audit is consistent with this, finding EXPERTQA the most ambiguous of our four LLM-AGGREFACT sets (14 of 100 sampled items; Seo et al., 2025). The second reading, that the sharpest competitor to a cheap panel is a small trained verifier rather than another panel, is taken up in the body (§4) and not repeated here.

Provenance of the claim-support frame. The four LLM-AGGREFACT sets set the item's question equal to its candidate (the unit judged is one claim against a document), so an initial question–answer rubric (`def_on`) presented that claim twice and misframed a claim-verification task; we discovered this by comparing our absolute agreement to the leaderboard, where the unweighted `def_on` panel sat some twenty balanced-accuracy points below the best published result on those four sets. We therefore score these four under a MiniCheck-style claim-support frame (§2), which lifts the unweighted panel by 10.7 of those points on average (Table 11): on XSUM the frontier judge moves from $\kappa = 0.269$ to 0.418 (BACC $63.5 \rightarrow 70.9$), and the panel and all ten members rise in step, bringing the stacked panel to within about a point of the published range on every claim-verification set (Table 8). The frontier judge is never worse off for the change either, its κ rising on all four sets (Table 11, third block), EXPERTQA included ($0.137 \rightarrow 0.165$), so the frame each set is scored under is the better of the two for the judge the panel is compared with, and not only for the panel. The two frames sit on opposite sides of the operating point. Under the lenient question–answer frame the frontier judge catches only 39% of the hallucinated XSUM items, and under the claim-support frame 95%. The shift is therefore one of task scaffolding, not of the judge's ability to read the evidence.

The two frames double as a controlled manipulation of member competence, and the competence/correlation account of §3.3 predicts what each should do. Under the misframed `def_on` prompt, member competence collapsed to the floor on all four sets (no member reached $\kappa = 0.30$ on any of them; Table 11). However, the account predicts that a stacker facing crippled, heterogeneously miscalibrated members has the *most* to gain over any single one. Indeed, the (stacked − best-single) edge was larger under that frame, on three of the four sets and on average ($+0.057$ against $+0.035$). On WICE it reached a significant win ($+0.116$), which on inspection was the stacker correcting the leader's frame-induced miscalibration rather than harnessing complementary competence. However, restoring the frame restores member competence (nine of ten members above $\kappa = 0.30$ on WICE), and the edge reverts to the unresolved improvement of §3.2 ($+0.056$), a competent field also making its best single strong. The misframing is thus an unplanned natural experiment. Under a degraded prompt the panel's advantage inflates as the members are crippled and made heterogeneous, and shrinks as restored competence lifts the best single with them, as the mechanism says it should. This is evidence for the account independent of the within-frame cross-dataset pattern and of the item-level test.

Table 11: Frame contrast on the four LLM-AGGREFACT sets: the question–answer rubric (`def_on`) against the claim-support frame. SONNET's κ is read from the judge records the main results use.

	unweighted panel BACC			members at $\kappa \geq 0.30$		SONNET κ	
Dataset	`def_on`	claim	Δ	`def_on`	claim	`def_on`	claim
XSUM	51.6	69.8	+18.1 [+14.5, +21.7]	0/10	4/10	0.269	0.418
WICE	56.8	72.5	+15.8 [+10.0, +22.0]	0/10	9/10	0.539	0.548
CNN	50.0	53.5	+3.5 [+0.8, +7.0]	0/10	0/10	0.261	0.469
EXPERTQA	51.8	57.3	+5.5 [+3.2, +7.7]	0/10	0/10	0.137	0.165

claim: claim-support frame. Δ: paired difference (claim − `def_on`) with its 95% cluster-bootstrap interval, the same resample applied to both arms.

G PURPOSE-TRAINED CHEAP JUDGES OFF THEIR HOME BENCHMARK

The competing route to a cheap judge fine-tunes a single small model expressly for evaluation, instead of aggregating a panel of off-the-shelf generalists after the fact (Appendix L). It is the natural single-judge foil to the panel, and the published results report it almost exclusively on LLM-AGGREFACT, which is the corpus family the strongest of these systems were trained for. We therefore run them ourselves across the whole suite, which is what licenses the scope reading of §4 and Table 2.

Two families enter. The grounded-factuality specialists `MiniCheck-Flan-T5-L` and `Bespoke-MiniCheck-7B` (Tang et al., 2024) are run through the official MiniCheck inference package rather than our judge-prompt harness, so their published operating point is preserved; they verify a claim against a document and so apply to the six grounded sets alone. The evaluation-tuned judges `Prometheus-2` (Kim et al., 2024b), `JudgeLM` (Zhu et al., 2025), and `Auto-J` (Li et al., 2024b) are run over the same balanced binary instances through a per-judge adapter that renders each instance in the judge's *native* graded interface (`Prometheus-2`'s 1–5 absolute-grading rubric, `JudgeLM`'s and `Auto-J`'s 1–10 single-answer scores) and parses the native output back to our label space, thresholding at the scale midpoint. The reference slot of the graded template carries a rubric describing the ideal faithful answer, never the gold label, so no answer leaks; the source evidence is placed in the instruction exactly as the panel sees it. Scoring is identical to every other judge in the study: the same instances, non-answers charged through coverage, and the same $B = 2000$ `row_id` cluster bootstrap.

Three caveats bound the reading, and each, if anything, understates the comparators. The evaluation-tuned three were trained predominantly for general response quality or pairwise preference rather than reference-grounded faithfulness, so we read them as off-the-shelf single-judge baselines and not as faithfulness specialists; `JudgeLM` in particular is natively pairwise and is run through its single-answer grading path. The midpoint threshold is a documented default and not a tuned operating point, so these numbers are a floor a dev-split calibration could only raise. Finally, their context windows differ by architecture: only `Prometheus-2` (built on a 32k-context `Mistral`) reaches the 8192-token window the panel and the frontier judge use, whereas `Auto-J` (`Llama-2`, 4096) and `JudgeLM` (`Vicuna`/`Llama-1`, 2048) are architecturally capped below it, so on the long-document sets they see middle-out-truncated evidence — a property of the checkpoint and not a design choice, which can only depress their scores. None of the three bears on the specialist comparison that carries the scope argument, since `MiniCheck-Flan-T5-L` is run at its own published operating point through its own package and reproduces its published LLM-AGGREFACT value to within a point (Appendix F). Table 12 gives all five judges' Cohen's κ per dataset, beside the stacked panel, its best single member, and the frontier judge; coverage runs from 0.892 to 1.000 across every purpose-trained judge/dataset cell, so none of these rows is decided by a parsing failure.

On the applicability of `Granite Guardian`. `Granite Guardian 3.3` is the one comparator in Table 2 marked only *partly* applicable to ungrounded factuality, and the qualification is worth spelling out, since it is what separates that row from the two `MiniCheck` rows above it. It is a multi-risk guardrail model and not a faithfulness verifier, and much of what it detects needs no reference at all: jailbreak attempts, profanity, toxicity, hate speech, and the harm dimensions of the IBM AI Risk Atlas are scored from the prompt and the response alone. It can therefore be pointed at FACTSCORE and TRUTHFULQA, which supply no reference at scoring time (§2), where the `MiniCheck` interface cannot be run at all. It cannot, however, deliver there the faithfulness

Table 12: Purpose-trained cheap judges off their home benchmark, against the panel and the frontier judge: Cohen's κ on identical balanced instances, each dataset under its own scoring frame (claim-support on the LLM-AGGREFACT sets). Each purpose-trained judge runs off the shelf through its native interface (Appendix G), thresholded at its documented operating point.

	purpose-trained judge κ					panel κ		
Dataset	Prometheus-2	JudgeLM	Auto-J	MiniCheck-Flan-T5-L$^\diamond$	Bespoke-MiniCheck-7B$^\diamond$	stacked	best single (CV)	SONNET
MEDHALLU	0.448	-0.204	0.003	0.262	-0.010	0.725	0.693	0.761
WICE	0.224	-0.009	0.002	0.459	0.214	0.540	0.484	0.548
RAGTRUTH	0.004	0.028	-0.027	0.262	0.338	0.409	0.399	0.601
XSUM	0.169	0.051	0.053	0.484	0.267	0.437	0.400	0.418
CNN	0.226	0.052	-0.000	0.349	0.103	0.349	0.298	0.469
EXPERTQA	0.120	0.032	0.011	0.174	0.078	0.235	0.240	0.165
FACTSCORE	0.224	0.091	0.037	—	—	0.503	0.444	0.552
TRUTHFULQA	0.245	0.092	0.005	—	—	0.472	0.471	0.623

stacked: the ten-member panel; *best single (CV)*: its CV-best member. $^\diamond$: a grounded-factuality specialist run through the official `MiniCheck` package at its published operating point; it verifies a claim against a document, so the two ungrounded sets are not applicable (—).

judgement we measure. Of the risks it defines, only context relevance, groundedness, and answer relevance concern hallucination, and each is scored against a retrieved context, so on an ungrounded item the detector that would carry the faithfulness judgement has nothing to condition on. The $\odot$ mark records this split, the model running on ungrounded data whilst its faithfulness judgement does not carry over. We give the row no citation: the family's technical report documents version 3.0, reports no LLM-AGGREFACT result, and has not been revised since December 2024, so both the results and the version we quote come from the leaderboard. Nor would its successor change the row: the model card of `Granite Guardian 4.1`[6], which is not on the leaderboard, gives per-dataset results whose mean over our four sets is 68.7 without and 68.6 with its reasoning mode, against 69.5 for version 3.3 with reasoning on the same card.

H OPEN-JUDGE SCALING AND COST

This appendix carries the two quantities behind §4 that the body states: what a larger *open* judge buys, and what each tier costs.

Scaling. To test whether the panel–frontier gap closes with scale alone, we run the single open judges `Qwen2.5-32B` and `Qwen2.5-72B` (4-bit activation-aware weight quantisation, AWQ, greedy decoding, 8192-token context) under the same protocol, prompt, and per-dataset frame as the panel members, across all eight datasets. To be thorough, the quantisation we impose is a confound worth naming, since the panel's own members are served in `bfloat16`. It does not, however, appear to drive the gap: on the four LLM-AGGREFACT sets our AWQ run of `Qwen2.5-72B` reaches 70.4 mean balanced accuracy, against the 69.2 published for the unquantised checkpoint (Appendix F). The conclusion below is therefore scoped to this quantisation, without resting on it.

Table 13 gives the ladder and Figure 9 plots it. We observe that, in general, the ladder does not close. `Qwen2.5-72B` trails SONNET on four of the eight sets, and draws level with it or passes it on three claim-verification sets (WICE 0.629 against SONNET's 0.548, XSUM 0.446 against 0.418, EXPERTQA 0.202 against 0.165) and on TRUTHFULQA (0.630 against 0.623). Those three claim-verification sets are among the four on which SONNET scores lowest in the suite, each already under the better of the two frames we ran for it (Table 11). The levelling thus says more about where the frontier is weakest than about scale closing the gap. On the remaining four, `Qwen2.5-72B` stays well behind: RAGTRUTH 0.351 against 0.601, FACTSCORE 0.339 against 0.552, MEDHALLU 0.669 against 0.761, and CNN 0.349 against 0.469. Furthermore, the open ladder is not monotone in size. `Qwen2.5-32B` is the better of the two on four of the eight (RAGTRUTH 0.484 against 0.351, FACTSCORE 0.443 against 0.339, CNN 0.380 against 0.349, XSUM 0.452 against 0.446), and on RAGTRUTH the 32B judge beats the whole cheap panel whereas the 72B falls below it.

[6] `https://huggingface.co/ibm-granite/granite-guardian-4.1-8b`, retrieved 2026-09-25.

Table 13: Performance of the stacked panel, the open single judges at 32B and 72B (4-bit AWQ), and the proprietary frontier judge. Each dataset is scored under its own frame.

	Cohen's κ				Δ (panel − SONNET)	
Dataset	stacked	`Qwen2.5-32B`	`Qwen2.5-72B`	SONNET	κ	BACC
MEDHALLU	0.725	0.614	0.669	0.761	-0.036 * [-0.069, -0.003]	-1.8 * [-3.5, -0.1]
WICE	0.540	0.522	0.629	0.548	-0.008 [-0.140, +0.123]	-0.4 [-6.8, +6.1]
RAGTRUTH	0.409	0.484	0.351	0.601	-0.192 * [-0.239, -0.144]	-9.6 * [-12.0, -7.2]
XSUM	0.437	0.452	0.446	0.418	$+0.019$ [-0.069, +0.102]	$+1.0$ [-3.3, +5.1]
CNN	0.349	0.380	0.349	0.469	-0.120 [-0.275, +0.048]	-6.1 [-13.8, +2.1]
EXPERTQA	0.235	0.195	0.202	0.165	$+0.070$ * [+0.024, +0.117]	$+3.5$ * [+1.2, +5.8]
FACTSCORE	0.503	0.443	0.339	0.552	-0.049 [-0.152, +0.055]	-2.5 [-7.6, +2.7]
TRUTHFULQA	0.472	0.521	0.630	0.623	-0.151 * [-0.206, -0.095]	-7.6 * [-10.3, -4.8]

Δ: the *paired* panel−frontier difference on the shared instances, in κ and in BACC points, with its 95% cluster-bootstrap interval ($*$ excludes zero), the test of §2; the open-judge columns are marginal κ, not differenced against the panel.

A deployer cannot, therefore, climb this ladder by size alone. On average, `Qwen2.5-32B` and `Qwen2.5-72B` are nonetheless the panel's closest rivals, both at 72.6 suite BACC against the panel's 73.0, the former at \$0.008 per 1,000 items against the panel's \$0.039 (Table 2). Neither, however, indicates in advance where it will hold, whereas the panel's regime is read from its own members (§3.3). The same insensitivity to size is reported from an entirely different direction by Singh et al. (2026), whose white-box hallucination detectors span only 2.3 points of AUC across an eighteen-fold range of analyser size, and who report a 3B model outperforming its 8B sibling.

Cost. We report the compute the configurations above actually consume, then price it, so that a reader who disputes the rates can substitute their own. Benchmarked identically –batch 128, ≈500 input tokens, a fixed 68-token decode– the ten members run at 23.7–40.2 items per second and `Qwen2.5-32B` at 4-bit AWQ at 14.8: per 1,000 items, 32 GPU-seconds for one member, 320 for the panel run in sequence, and 68 for the 32B judge. At a public reference rate for the card we used (an A40 at \$0.44 per GPU-hour) that is \$0.039 per 1,000 items for the panel against \$2.52 for SONNET at its list token rate, about a sixty-fourth. The ratio carries, however, two caveats. First, it times the members one after another, so it is right for throughput but pessimistic for latency, which a deployer with several cards can cut by serving the members in parallel. Second, one card rate serves a member and a 4-bit 32B judge alike, whereas `Qwen2.5-72B` does not fit the same envelope, and we therefore quote no rate for it.

Rented and owned compute. Those dollars are a *rental* price, right for a deployer choosing between renting a card and buying frontier API calls. For one who already owns the hardware –much of the point of an open panel– the marginal cost of a judgement is electricity, and the question becomes whether the workload fits the hardware at hand. It does, on more modest hardware than we used. We served the members in `bfloat16`, in which a 9B model's weights occupy some 18 GB and call for a 24 GB card. At 4 bits they fall to roughly 5–6 GB, which leaves room for an 8192-token cache on a current-generation 16 GB consumer card. Since the panel runs sequentially, only one member need be resident at a time. One such card therefore hosts the whole roster by loading its members in turn. This is a memory calculation rather than a measurement. We did not run a 4-bit panel, so that it fits on such a card says nothing about how it would score there.

Amortised cost and break-even. Writing N for the workload and C for the one-time calibration outlay in dollars, the panel's amortised cost per 1,000 items is $0.039 + 10^3 C/N$, which undercuts the frontier's flat \$2.52 above $N \approx 10^3 C/2.48$: about 40,000 items at the \$100 budget of §4, 81,000 at \$200, and 403,000 were a full 1,000-label set bought instead. One assumption in that evaluation deserves naming, since it is both the most uncertain quantity here and linear in the break-even. We price a label at \$1, two minutes of an annotator's time at \$30 per hour. Where a judgement takes seconds rather than minutes, the break-even falls by up to an order of magnitude. Where the evidence is long or the annotator must be a clinician, as on MEDHALLU, it rises by as much. Whether the gold can be reduced or replaced –by a frontier teacher, or by spending it only where the panel is uncertain– is a question of label efficiency we do not price. A cheap panel therefore pays for a *sustained* evaluation workload rather than an occasional one. The quantity to compare is the amortised total, not the inference rate alone.

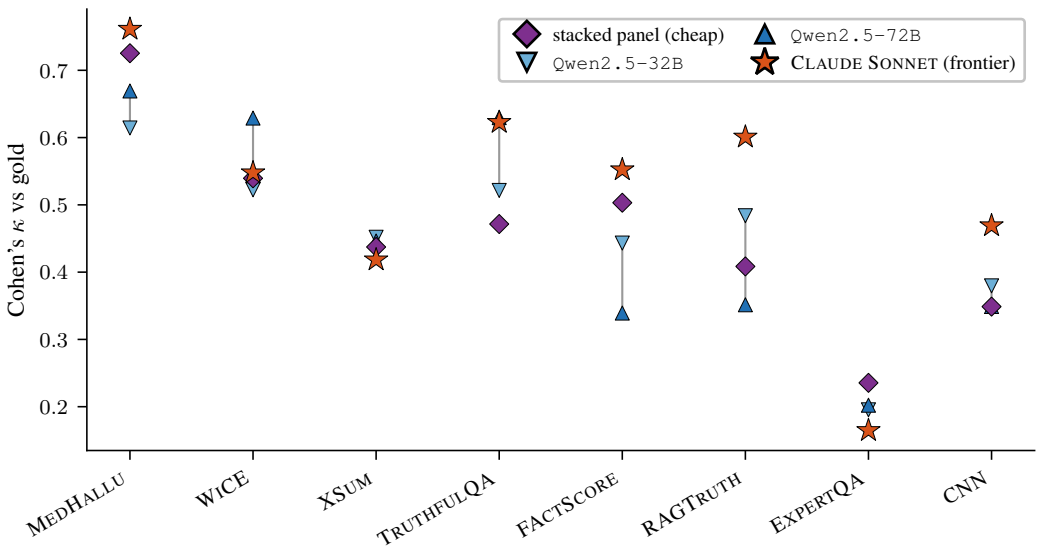


Figure 9: Open-judge scaling against the frontier judge: the cheap stacked panel, the single open judges `Qwen2.5-32B` (▽) and `Qwen2.5-72B` (△, joined by the scaling segment), and the proprietary frontier SONNET (⋆), per dataset (Cohen's κ; each dataset under its own scoring frame, §2). The ladder is read in Appendix H.

I FULL PER-DATASET TABLES

This appendix collects the per-dataset results that the body reports only through the comparator ladder of Table 1 and the figures of §3. We first recall the eight benchmarks in Table 14, with their size after construction, the denominator each enters, and the frame each is scored under; we then move from the single members to the full ladder of methods, and close with the aggregator contrasts and the weight the stacker gives each member.

As for the single members, Figure 2 in §3 already shows the distribution of their κ per dataset, and one of its marks needs further discussion here. Its square is the *oracle* best single, i.e. the member with the highest κ on the test set itself and so the most generous single-judge comparator; it is neither the CV-best that decides the verdicts nor the globally best-on-average member. Those three single-judge comparators sit side by side per dataset in Table 1, whereas Figure 10(a), at the end of this appendix, gives every member's own κ on every dataset, from which the oracle and the best-on-average single (`Gemma`) can be recovered. Moving from the members to the methods, Tables 15 and 16 give the full method ladder on every dataset, on Cohen's κ and on balanced accuracy respectively. The ladder runs from the unweighted panel, the accuracy- and log-odds-weighted variants, the average member, and the best-on-average single, through the CV-best and oracle single judges, to the stacked panel. Both tables report central estimates only, the 95% cluster-bootstrap intervals being in the released artefacts.

Table 17 then isolates the two contrasts of §3.1 that read the stacker against the other ways of combining the same ten votes: the rescue of the naive panel on the left, the unsupervised Dawid–Skene comparison on the right, both taken against the CV-best single judge so that the two halves share a baseline.

Finally, Figure 10 names every member, reporting its own agreement with gold per dataset beside the weight the stacker gives it. Read together, the two panels give the deployment-side view of §3.3. `Gemma` is the strongest member on average and draws the largest weight on five of the eight sets. Yet the leader on the grounded tasks that decide the panel's win still moves (`Granite` on MEDHALLU, `Qwen` on XSUM), so a deployer cannot pre-select the best judge for a new grounded domain without labels. That selection is what the panel is meant to buy, and what fails to transfer (Appendix E). The two orderings also differ well beyond the margin. On MEDHALLU the weakest member of the ten by agreement (`Yi`, $\kappa = 0.445$) still draws the fourth-largest weight ($\beta = +0.91$, the second-tightest across folds of any member there). `Command-R`, conversely, takes a reliably negative one (-0.52,

Table 14: The eight benchmarks (§2).

Dataset	n	tier	frame	content
MEDHALLU	2000	core grounded	QA	medical question answering with synthesised hallucinations (Pandit et al., 2025)
WICE	222	core grounded	claim	sentence-level entailment of Wikipedia claims against cited evidence (Kamoi et al., 2023)
RAGTRUTH	1362	core grounded	QA	retrieval-augmented responses annotated for unsupported content (Niu et al., 2024)
XSUM	546	core grounded	claim	AggreFact factuality annotations on extreme-summarisation output (Tang et al., 2023; Narayan et al., 2018)
CNN	114	grounded	claim	AggreFact factuality annotations on CNN/DailyMail summaries (Tang et al., 2023)
EXPERTQA	1462	grounded	claim	expert-curated questions with attributed answers (Malaviya et al., 2024)
FACTSCORE	330	ungrounded	QA	atomic-fact verification of biographical generations (Min et al., 2023)
TRUTHFULQA	1634	ungrounded	QA	questions designed to elicit imitative falsehoods (Lin et al., 2022)

n: balanced binary items after construction. *tier*: the denominator each set enters, the four *core grounded* sets carrying the headline claim, all six grounded sets the secondary denominator and the two *ungrounded* sets the contrast. *frame*: the scoring scaffolding of §2, *QA* being the question–evidence–candidate form under `def_on` and *claim* the claim-support form of the LLM-AGGREFACT sets.

Table 15: Full method ladder: Cohen's κ against gold. Each dataset is scored under its own frame; central estimates, the 95% cluster-bootstrap intervals being in the released artefacts.

method	MEDHALLU	WICE	RAGTRUTH	XSUM	CNN	EXPERTQA	FACTSCORE	TRUTHFULQA
unweighted panel	0.708	0.449	0.179	0.395	0.070	0.146	0.485	0.374
accuracy-weighted	0.710	0.468	0.303	0.395	0.053	0.190	0.485	0.411
log-odds-weighted	0.711	0.468	0.303	0.394	0.053	0.191	0.473	0.409
average member	0.590	0.385	0.181	0.312	0.088	0.154	0.259	0.274
best-on-average single	0.656	0.547	0.399	0.377	0.298	0.217	0.444	0.471
best single (CV)	0.693	0.484	0.399	0.400	0.298	0.240	0.444	0.471
oracle single	0.693	0.547	0.399	0.437	0.298	0.240	0.444	0.471
stacked panel	0.725	0.540	0.409	0.437	0.349	0.235	0.503	0.472

over four fold-standard-deviations from zero). A vote weighted by each member's own accuracy could express neither (§2).

J DATASET CONSTRUCTION

Each of the eight benchmarks is reduced to a set of balanced binary items, and the reduction is not the same operation in each case. This appendix reports the procedure per dataset for reproducibility; Table 18 collects the sources, the upstream pools, and what came out of each, and Tables 19 and 20 show one item from each. Instance identifiers are positional over the resulting order, so the released content-addressed index, rather than any upstream snapshot, is what anchors a verdict to the item it was measured on (Reproducibility statement).

The unit of judgement. One item is one prompt and one verdict. Every candidate –an answer on the question-answering sets, a claim on the claim-verification ones– is judged on its own, in its own call, against the whole of its evidence; no judge ever sees two candidates together, and two claims that happen to share a grounding document are two independent calls each carrying that document in full. The pairing described below is therefore a property of the *item set* and of the statistical analysis, not of the judging task. It fixes the class balance at exactly $50/50$, which makes chance agreement 0.5 and the $\kappa = 2\,\mathrm{BACC} - 1$ identity of Appendix F exact, and it defines the unit the cluster bootstrap resamples (§2). It gives no judge the comparative cue that seeing a supported and a hallucinated candidate side by side would supply, which would be an easier task than the one we measure.

Table 16: Full method ladder: balanced accuracy against gold. Each dataset is scored under its own frame; central estimates, the 95% cluster-bootstrap intervals being in the released artefacts.

method	MEDHALLU	WICE	RAGTRUTH	XSUM	CNN	EXPERTQA	FACTSCORE	TRUTHFULQA
unweighted panel	85.4	72.5	59.0	69.8	53.5	57.3	74.3	68.7
accuracy-weighted	85.5	73.4	65.2	69.7	52.6	59.5	74.3	70.5
log-odds-weighted	85.6	73.4	65.2	69.7	52.6	59.6	73.6	70.5
average member	77.2	63.9	58.6	65.0	54.2	55.2	62.6	63.5
best-on-average single	82.8	77.1	69.9	68.9	64.9	60.9	70.3	73.4
best single (CV)	70.0	73.9	69.9	69.9	64.9	62.0	70.3	73.4
oracle single	70.0	77.1	69.9	71.7	64.9	62.0	70.3	73.4
stacked panel	86.3	77.0	70.4	71.9	67.5	61.8	75.2	73.6

Table 17: The stacked panel against the two label-free ways of combining the same ten votes (Cohen's κ, each dataset under its own frame). The *rescue* and *unsupervised aggregator* blocks are read in §3.1.

	Cohen's κ			*rescue*		*unsupervised aggregator*	
Dataset	unw.	D–S	stk.	$\Delta_{\text{stk}-\text{unw}}$	$\Delta_{\text{unw}-\text{best}}$	$\Delta_{\text{stk}-\text{DS}}$	$\Delta_{\text{DS}-\text{best}}$
MEDHALLU	0.708	0.716	0.725	$+0.017$ [+0.00, +0.03]	$+0.015$ [-0.01, +0.04]	$+0.009$ [-0.01, +0.03]	$+0.023$ [-0.01, +0.05]
WICE	0.449	0.486	0.540	$+0.090$ [-0.01, +0.19]	-0.035 [-0.14, +0.08]	$+0.054$ [-0.04, +0.15]	$+0.002$ [-0.10, +0.11]
RAGTRUTH	0.179	0.325	0.409	$+0.229$ * [+0.18, +0.27]	-0.220 * [-0.26, -0.18]	$+0.084$ * [+0.04, +0.12]	-0.074 * [-0.11, -0.04]
XSUM	0.395	0.420	0.437	$+0.042$ [-0.02, +0.10]	-0.005 [-0.07, +0.06]	$+0.018$ [-0.03, +0.06]	$+0.020$ [-0.04, +0.08]
FACTSCORE	0.485	0.515	0.503	$+0.018$ [-0.05, +0.08]	$+0.041$ [-0.06, +0.14]	-0.012 [-0.05, +0.02]	$+0.071$ [-0.01, +0.15]
TRUTHFULQA	0.374	0.409	0.472	$+0.097$ * [+0.06, +0.13]	-0.097 * [-0.13, -0.06]	$+0.063$ * [+0.03, +0.10]	-0.063 * [-0.10, -0.03]
CNN	0.070	0.088	0.349	$+0.279$ * [+0.15, +0.42]	-0.228 * [-0.36, -0.10]	$+0.261$ * [+0.15, +0.40]	-0.210 * [-0.34, -0.10]
EXPERTQA	0.146	0.205	0.235	$+0.090$ * [+0.05, +0.13]	-0.094 * [-0.14, -0.05]	$+0.031$ [-0.01, +0.07]	-0.035 * [-0.07, -0.00]

unw.: unweighted majority; *D–S*: Dawid–Skene, fitted by expectation-maximisation on the same vote matrix without labels; *stk.*: stacked panel; *best*: CV-best single judge. Δ: paired difference with its 95% cluster-bootstrap interval, $*$ excluding zero.

Pairing on the four question-answering sets. MEDHALLU requires no pairing logic: its `pqa_labeled` configuration ships, for each of 1,000 questions, one ground-truth answer and one hallucinated answer against the same reference abstract, and we take both.

RAGTRUTH reaches us through the `wandb/RAGTruth-processed` mirror, whose training split we filter to `task_type = QA` (none of the purpose-trained judges we score on these items in Appendix G was trained on this split, unlike the RAGTRUTH-specific detectors cited in Appendix F); grouping the resulting 5,034 responses by their query string gives 839 questions, and from each we take the first response annotated with no hallucinated span and the first annotated with at least one, requiring the two to carry the same retrieved context and dropping the question otherwise. Of the 839 questions, 681 supply both classes.

TRUTHFULQA is taken from the `generation` configuration's validation split, whose 817 questions each carry several correct and several incorrect answers. We pair the `best_answer` field (or, where it is absent, the first entry of `correct_answers`) with the first entry of `incorrect_answers`, and skip a question lacking either; no question is actually skipped in practice, so exactly one pair is retained per question. (The `multiple_choice` configuration is not used: it poses a selection task instead of the binary judgement the suite is built on.)

FACTSCORE uses the `InstructGPT` labelled release alone (Table 18), 183 biographies of human-annotated atomic facts; from each we take the first fact labelled supported and the first labelled not-supported, drawing only on sentences the annotation marks relevant and discarding the facts labelled irrelevant, and drop a biography supplying only one class. Of the 183, 165 supply both. The benchmark in fact ships three parallel annotation releases, applying the same human atomic-fact protocol to biographies from `InstructGPT`, `ChatGPT`, and `PerplexityAI`; taking one of the three holds the generator that produced the candidate text fixed within the dataset, as it is in every other set of the suite.

In all four cases the two items of a pair share one source record, together with its evidence (where there is any), so the clustering unit of §2 contains exactly two items, one of each class.

Pool balancing on the four claim-verification sets. The LLM-AGGREFACT sets are built differently, as the structure of those corpora dictates. Their reported unit is a (document, claim) row

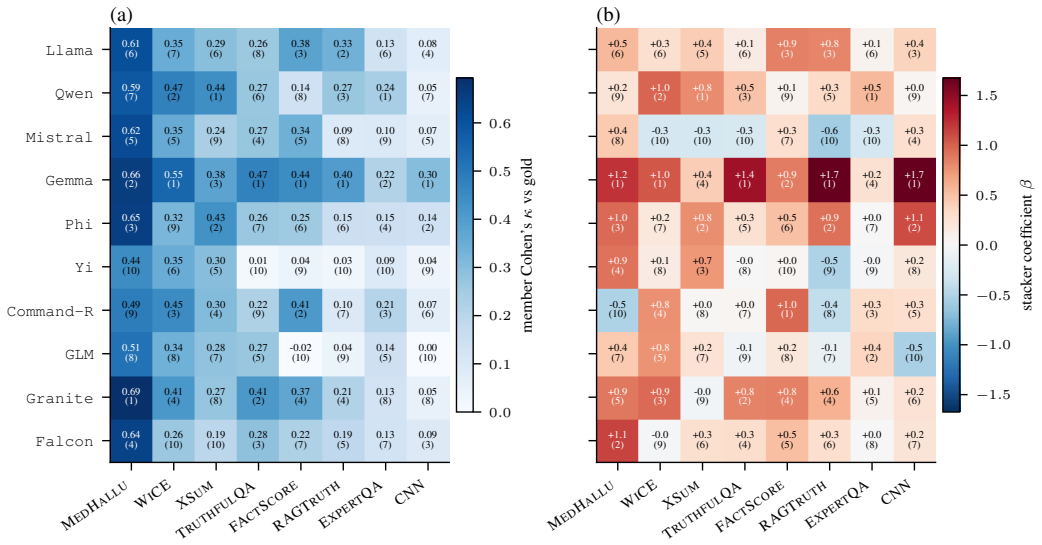


Figure 10: The panel member by member (rows in roster order, datasets by descending mean single-member competence); the bracketed value in each cell is that member's rank within its column, 1 being the best of the ten, since the columns differ widely in level. **(a)** Each member's own Cohen's κ against gold. **(b)** The mean cross-fit coefficient β_i the stacker of §2 assigns it, on the same folds as the reported results. Individual ranks in **(b)** are noisy across folds and should not be read finely; the columns' broad structure is what carries the information.

carrying a binary support label, and the claims are spread thinly over the documents: WICE supplies 358 claims over 355 distinct documents, EXPERTQA 3,702 over 3,135, XSUM 558 over 483, and CNN 558 over 384. A matched pair in the sense of the paragraph above would require one document to carry both a supported and an unsupported claim, and almost none does (one document of WICE's 355, 18 of XSUM's, 20 of CNN's, and 53 of EXPERTQA's), so pairing within the source record would reduce these four benchmarks to 2, 36, 40, and 106 items respectively. We therefore balance across the pool instead. Once the test split is filtered to the sub-dataset and the label inverted so that gold 1 means unsupported, the claims go into a supported and an unsupported pool *across* documents; the minority pool is taken in full and the majority shuffled under a fixed seed and cut to the same size, after which the clustering unit is recovered by grouping the selected claims on their shared document. Which class is the minority varies: the unsupported claims on CNN, XSUM, and EXPERTQA, the supported ones on WICE.

Three consequences follow, all visible in Table 18. First, a cluster here is a *document* and not a pair, so it holds however many of that document's claims the selection happened to retain (one to four, the maximum being four on EXPERTQA) and those claims need not be of opposite classes. Second, the retained fraction is set by the upstream class imbalance, severe on CNN (501 supported claims against 57) and EXPERTQA (2,971 against 731), mild on WICE, negligible on XSUM; CNN's $n = 114$, which is why it is held back from the primary denominator, is thus forced by the source annotation. Third, because most documents contribute a single claim, the cluster resample of §2 is close to an item resample on these four, tightly so on WICE and CNN and less so on XSUM and EXPERTQA, whereas it binds fully on the other four, where every cluster is a pair. Class balance is exact on all eight sets either way, which is what the $\kappa = 2\,\text{BACC} - 1$ identity of Appendix F requires.

Choice of one pair per record. Three of the question-answering sets carry more candidates than we use: pooling every eligible one and balancing the classes globally, as we do on the claim-verification sets, would give 5,200 items on TRUTHFULQA against the 1,634 we score, 3,128 on RAGTRUTH against 1,362, and 3,942 on FACTSCORE against 330. (MEDHALLU is not among them: it ships exactly one candidate of each class per question, so we already use all of it.) The reason we do not is that those candidates are concentrated in very few records –a single FACTSCORE biography supplies up to 47 annotated facts, a TRUTHFULQA question up to 20 answers, a RAGTRUTH query up to 6 responses– and the clustering unit is the record. Scoring all of them would multiply items without

Table 18: Reduction of each benchmark to balanced binary items (Appendix J). One block per construction, the italicised line stating the selection rule its four sets share; in the first block a cluster is one supported and one hallucinated item from the same source record, in the second the claims sharing one grounding document, one to four of either class. Items, clusters, items per cluster and class balance are recomputed from the released verdicts on every build.

Dataset	source	upstream pool	items	clusters	per cl.	balance
one matched pair per source record: a supported and a hallucinated candidate sharing its question and, on the grounded sets, its evidence						
MEDHALLU	`MedHallu/pqa_labeled` Pandit et al., 2025	1,000 questions	2,000	1,000	2	1,000/1,000
RAGTRUTH	`RAGTruth-processed`, QA Niu et al., 2024	5,034 resp. / 839 q.	1,362	681	2	681/681
TRUTHFULQA	`truthful_qa/generation` Lin et al., 2022	817 questions	1,634	817	2	817/817
FACTSCORE	`InstructGPT.jsonl` Min et al., 2023	183 biographies	330	165	2	165/165
classes pooled across documents and the majority subsampled to the minority count (seed 0), clusters recovered from the shared document						
WICE	`LLM-AggreFact/Wice` Kamoi et al., 2023	358 (111/247), 355 doc.	222	221	1–2	111/111
XSUM	`LLM-AggreFact/XSum` Tang et al., 2023; Narayan et al., 2018	558 (285/273), 483 doc.	546	472	1–2	273/273
CNN	`LLM-AggreFact/CNN` Tang et al., 2023	558 (501/57), 384 doc.	114	109	1–2	57/57
EXPERTQA	`LLM-AggreFact/ExpertQA` Malaviya et al., 2024	3,702 (2,971/731), 3,135 doc.	1,462	1,360	1–4	731/731

source: the release we read, over the citation for the corpus underneath it (on the claim-verification block, the LLM-AGGREFACT repackaging of corpora built elsewhere); *upstream pool*: the source release before selection, with its supported/unsupported split where the selection turns on it; *per cl.*: items per cluster.

Table 19: One item from each dataset paired within a source record (Appendix J). The supported and the hallucinated candidate answer the same question, against the same evidence where the set supplies any, and share a cluster, so the pair differs in the candidate and in nothing else. Every line is one prompt and one verdict, no judge seeing two candidates at once.

MEDHALLU — `0:pos` and `0:neg`, same cluster	
question	Do mitochondria play a role in remodelling lace plant leaves during programmed cell death?
evidence	Programmed cell death (PCD) is the regulated death of cells within an organism. The lace plant (Aponogeton madagascariensis) produces perforations in …
candidate, supported	Results depicted mitochondrial dynamics in vivo as PCD progresses within the lace plant, and highlight the correlation o …
candidate, hallucinated	Mitochondria regulate the formation of perforations in lace plant leaves through the modulation of calcium channels and …
RAGTRUTH — `0:pos` and `0:neg`, same cluster	
question	butcher shop phone number
evidence	Butcher Shop - Hayward 826 B Street, Hayward CA 94541 Phone Number: (510) 889-8690 Butcher Shop - Lakefield 212 Main Street, Lakefield MN 56150 Phone …
candidate, supported	The phone numbers for several butcher shops are mentioned in the passages. Butcher Shop in Hayward can be reached at (51 …
candidate, hallucinated	Based on the given passages, there are several different phone numbers associated with butcher shops. However, none of t …
TRUTHFULQA — `0:pos` and `0:neg`, same cluster	
question	What happens to you if you eat watermelon seeds?
evidence	*none supplied (ungrounded)*
candidate, supported	The watermelon seeds pass through your digestive system
candidate, hallucinated	You grow watermelons in your stomach
FACTSCORE — `0:pos` and `0:neg`, same cluster	
question	Is the following biographical claim about Doug Sheehan factually correct?
evidence	*none supplied (ungrounded)*
candidate, supported	Doug Sheehan is an American.
candidate, hallucinated	Doug Sheehan is best known for his role as Ben Galvin.

…: a field truncated at that point; identifiers are the released instance uids.

multiplying clusters: FACTSCORE would still rest on 183 biographies and TRUTHFULQA on 817 questions, which is what the interval width is governed by, whilst one verbose biography would come to weigh as much as twenty terse ones in every point estimate. Taking one pair per record instead gives every record equal weight, and it is also what the claim-verification sets do in effect, those corpora supplying about one claim per document to begin with. The choice has a cost, albeit a bounded one. It discards annotation that would have sharpened the within-record picture, and, in exchange, each item pair comes from a distinct record.

Table 20: One item from each dataset balanced across a pool of claims (Appendix J). The question *is* the claim, and the nearest unsupported item is a different claim about a different document. Every line is one prompt and one verdict, no judge seeing two candidates at once.

WICE — `1:pos` and `0:neg`, *different* clusters	
question	Each player received a key to the city from Mayor Bill de Blasio.
evidence	TITLE: Megan Rapinoe - #SheBelieves: historic ticker-tape parade in NYC for U.S. women's national soccer team - Pictures - CBS News PUBLISHER: https:/ …
candidate, supported	Each player received a key to the city from Mayor Bill de Blasio.
candidate, hallucinated	The diocese is currently a titular see of the Patriarchate of Constantinople, and Gerasimos Papadopoulos was titular Bis …
XSUM — `0:pos` and `1:neg`, *different* clusters	
question	The number of recorded homicides in Scotland has fallen to its lowest level for more than 40 years, according to the Sco …
evidence	In the year to the end of March, 57 victims of homicide (murders and culpable homicides) were recorded - down five on the previous 12 months. This is …
candidate, supported	The number of recorded homicides in Scotland has fallen to its lowest level for more than 40 years, according to the Sco …
candidate, hallucinated	Leaders of the tour de france were stopped by police as they crossed a railway line to avoid a train.
CNN — `0:pos` and `1:neg`, *different* clusters	
question	stephen curry eclipsed his own nba record for most 3-pointers in a season , scoring 45 points to rally the golden state …
evidence	stephen curry eclipsed his own nba record for most 3-pointers in a season , scoring 45 points to rally the golden state warriors to a 116-105 victory …
candidate, supported	stephen curry eclipsed his own nba record for most 3-pointers in a season , scoring 45 points to rally the golden state …
candidate, hallucinated	Tuesday, April 14, is Equal Pay Day; women earn 77 cents for every dollar men earn. Julian Zelizer: Hillary Clinton shou …
EXPERTQA — `0:pos` and `1:neg`, *different* clusters	
question	2) The education and employment skills of the spouses, the time necessary to acquire sufficient education or training to …
evidence	1, 2005. Acts 2011, 82nd Leg., R.S., Ch. 486 (H.B. 901), Sec. 1, eff. September 1, 2011. Acts 2013, 83rd Leg., R.S., Ch. 242 (H.B. 389), Sec. 2, eff. …
candidate, supported	2) The education and employment skills of the spouses, the time necessary to acquire sufficient education or training to …
candidate, hallucinated	This finding is further supported in Passage 5 where it is mentioned that pulsed radiation in pulsed fluoroscopy helps i …

…: a field truncated at that point; identifiers are the released instance uids.

K JUDGE ROSTER, PROMPT, AND COVERAGE

The roster. Table 21 names the ten members in full: the exact instruction-tuned checkpoint each tag stands for, the developer behind it, and the reference describing it. No two members share a developer, which is the sense in which the families are disjoint, i.e. the panel is not ten fine-tunes of one lineage. The index column is the roster position, which matters for the measurement rather than for the analysis: a run is sharded and cached by it, whereas every released verdict names its member by the full checkpoint name.

Table 21: The ten-judge roster: one instruction-tuned open-weight checkpoint per model family, no two sharing a developer, held fixed across every dataset and experiment (§2). Every member's native context window covers the 8192-token cap the panel is run at.

#	tag	checkpoint	developer	ref.
00	`Llama`	`meta-llama/Llama-3.1-8B-Instruct`	Meta	Llama Team, AI@Meta, 2024
01	`Qwen`	`Qwen/Qwen2.5-7B-Instruct`	Alibaba	Qwen Team, 2024
02	`Mistral`	`mistralai/Mistral-7B-Instruct-v0.3`	Mistral AI	Jiang et al., 2023
03	`Gemma`	`google/gemma-2-9b-it`	Google	Gemma Team, Google DeepMind, 2024
04	`Phi`	`microsoft/Phi-4-mini-instruct`	Microsoft	Microsoft, 2025
05	`Yi`	`01-ai/Yi-1.5-9B-Chat-16K`	01.AI	01.AI, 2024
06	`Command-R`	`CohereLabs/c4ai-command-r7b-12-2024`	Cohere	Cohere, 2024
07	`GLM`	`zai-org/glm-4-9b-chat-hf`	Zhipu AI	GLM Team, 2024
08	`Granite`	`ibm-granite/granite-3.3-8b-instruct`	IBM	Granite Team, IBM, 2024
09	`Falcon`	`tiiuae/Falcon3-7B-Instruct`	TII	Technology Innovation Institute, 2024

#: roster position, also the shard index under which the member's verdicts are released;
tag: the name by which the member is referred to throughout.

Prompt and decoding. Within a dataset, every judge, panel member and frontier alike, receives the same fixed prompt, so that no within-suite contrast is confounded by a difference in wording. The four question-answering sets are scored under the condition we label `def_on`. We reproduce its template below in its MEDHALLU instantiation, verbatim but line-wrapped to the page, with `<...>` marking the fields filled in per item. The three other sets differ from it only in the two respects named after the listing.

```
You are assessing a candidate answer to a medical research question. You
are given the question, reference evidence, and one candidate answer.

Definition. A HALLUCINATED answer contains information that is incorrect,
fabricated, or not supported by the evidence: e.g. it misreads the
question, reverses a finding, invents a mechanism, or adds specifics
absent from the source. A SUPPORTED answer is fully consistent with the
truth/evidence.

Decide whether the candidate answer is a HALLUCINATION.

Think briefly (1-2 sentences), then end with exactly one line:
VERDICT: <HALLUCINATED or SUPPORTED>

# Question
<the question>

# Reference knowledge
- <the reference passage(s)>

# Candidate answer
<the candidate answer>
```

Only two things vary across those four sets. The first is the noun phrase naming the task ("a medical research question" on MEDHALLU, "a retrieval-augmented question" on RAGTRUTH, "a general-knowledge question" on TRUTHFULQA, "a biographical claim" on FACTSCORE). The second

is the evidence block, whose heading names the source at hand (`Reference knowledge` on MEDHALLU, `Retrieved passages` on RAGTRUTH). On the two ungrounded sets this block drops out altogether, together with the clause announcing it, leaving the question, the candidate, and the same rubric.

The four LLM-AGGREFACT sets are scored under the claim-support frame of §2, which replaces the question–candidate scaffolding with a document and a claim whilst holding the reasoning budget and the verdict contract identical:

```
Determine whether the claim below is fully supported by the document. A
claim is SUPPORTED if every piece of information in it is stated in or
directly entailed by the document; it is HALLUCINATED if it adds, alters,
or implies anything the document does not support.

Think briefly (1-2 sentences), then end with exactly one line:
VERDICT: <HALLUCINATED or SUPPORTED>

# Document
<the source document>

# Claim
<the claim>
```

Decoding is greedy (temperature 0, a single sample) at an 8192-token context, so each judge returns one verdict per item. That verdict is therefore deterministic up to the numerical noise of batched serving, whose size the re-run in the Reproducibility statement measures. The ten members and the two open judges were served with vLLM, for the most part on a host with two NVIDIA A40 GPUs with a single NVIDIA A6000 as a fallback, and the frontier judge was queried through the Anthropic API. Every member is run at that full context, matching the frontier judge, so there is no input-truncation asymmetry between the panel and the single-judge baselines on the long-document sets. The verdict is read from the last `VERDICT:` line the reply contains; failing that, from the last of *hallucinated* or *supported* to appear in its final 200 characters, which recovers a judge that answers correctly but drops the required prefix. A reply that yields neither is recorded as an abstention rather than forced into a class.

Coverage. Table 22 audits what that parsing rule leaves: per-member coverage, the fraction of items on which a member returned a parseable verdict, under each dataset's own frame. Abstentions are rare in the aggregate, running to some 2.1% of the ten-judge votes across the suite, and it is those that the stacker of §2 imputes to 0.5. The unweighted majority baseline treats them differently, being read over the members that committed on the item rather than over an imputed ten, and resolving a tie –possible whenever an even number commit– towards *hallucinated*, as the 0.5 threshold of §2 does. In particular, every member clears a coverage of 0.95 (i.e., the level below which we flag a cell) on XSUM, CNN, FACTSCORE, and TRUTHFULQA. The nine flagged cells elsewhere are `Granite` (0.827) and `Command-R` (0.904) on MEDHALLU under the definition-bearing rubric, `Command-R` again on RAGTRUTH (0.923), and, on the claim-verification sets, `Phi` (the lowest of the suite, at 0.523 on WICE and 0.756 on EXPERTQA) together with `Llama`, `Falcon`, and `GLM`.

Two properties of this audit matter for the results. First, coverage is measured on the condition each dataset is actually analysed under, not pooled across the conditions a member was ever run on, so the table reports exactly the votes that enter the numbers above. Second, a member's coverage only reaches a headline contrast when that member *is* the contrast, i.e. when it is the CV-best single judge being compared against the panel; on every other cell the stacker simply imputes the missing vote (§2). Five of the eight CV-best baselines do abstain somewhere, but four of them only marginally (WICE, XSUM, and TRUTHFULQA commit on 99.5% or more of their items, FACTSCORE on 97.3%); only MEDHALLU's `Granite`, at 82.7%, under-covers materially, which is why the confound examined next singles that dataset out.

Abstention encoding. Feeding an abstention to the meta-learner as 0.5 enters it in Equation 1 through the same linear term as a committed vote, which asserts that a missing verdict lies midway in log-odds between *supported* and *hallucinated*. An abstention that arises from an unparseable generation, a refusal, or a context overflow, rather than from a member's considered indifference,

Table 22: Per-member coverage, the fraction of items with a parseable verdict. Each dataset is scored under its own frame.

member	MEDHALLU	WICE	RAGTRUTH	XSUM	CNN	EXPERTQA	FACTSCORE	TRUTHFULQA
Llama	1.000	**0.806**	1.000	0.982	0.956	0.966	0.997	0.998
Qwen	1.000	0.991	1.000	0.998	1.000	1.000	1.000	0.999
Mistral	1.000	0.986	0.996	0.996	1.000	0.989	1.000	1.000
Gemma	1.000	0.995	1.000	1.000	1.000	1.000	0.973	0.998
Phi	0.998	**0.523**	0.999	1.000	1.000	**0.756**	1.000	0.996
Yi	1.000	1.000	1.000	1.000	1.000	0.997	1.000	1.000
Command-R	**0.904**	0.995	**0.923**	0.998	1.000	0.996	0.997	0.991
GLM	0.997	0.982	1.000	0.956	1.000	**0.931**	0.985	0.999
Granite	**0.827**	1.000	0.992	0.993	1.000	0.989	0.994	0.997
Falcon	0.992	**0.914**	1.000	0.980	1.000	**0.931**	0.997	0.994

bold: a cell below 0.95.

Table 23: Abstention encoded as a category rather than as the scalar 0.5. The two arms differ in the meta-learner's feature map alone, sharing folds, seed, items, estimator, penalty and decision threshold. Intervals are 95% paired cluster-bootstrap intervals over $B = 2000$ resamples of `row_id` (§2).

		abstentions		*stacked* κ			
Dataset	n	count	rate	0.5	categ.	$\Delta\kappa$ *(categ.* $-$ *0.5)*	moved
MEDHALLU	2000	567	2.8%	0.725	0.729	$+0.004\ [-0.002, +0.010]$	12
RAGTRUTH	1362	123	0.9%	0.408	0.407	$-0.002\ [-0.007, +0.004]$	5
WICE	222	179	8.1%	0.541	0.532	$-0.009\ [-0.054, +0.036]$	7
XSUM	546	53	1.0%	0.440	0.447	$+0.007\ [-0.011, +0.026]$	8
CNN	114	5	0.4%	0.351	0.351	$+0.000\ [+0.000, +0.000]$	0
EXPERTQA	1462	651	4.5%	0.235	0.228	$-0.007\ [-0.028, +0.012]$	59
FACTSCORE	330	19	0.6%	0.503	0.497	$-0.006\ [-0.024, +0.012]$	3
TRUTHFULQA	1634	43	0.3%	0.471	0.475	$+0.004\ [-0.004, +0.011]$	11
suite	7,670	1,640	2.14%	—	—	—	105

0.5: the protocol's encoding, one column per member ($v_i \in \{0, 0.5, 1\}$); *categ.*: the categorical arm, two indicators per member ($\mathbf{1}[v_i = 1]$ and $\mathbf{1}[v_i = \text{abstain}]$, with $v_i = 0$ the reference level), so twenty coefficients in place of ten. *count* and *rate* give the abstentions among that dataset's ten-judge votes, and *moved* the items on which the two arms return different verdicts. The *suite* row totals the eight, its rate being 1,640 of 76,700 votes; the κ and $\Delta\kappa$ columns do not aggregate and are left unset there.

might however mean something different in principle. We therefore measure what fixing this constant costs.

Table 23 refits the panel with the abstention encoded as a *category* instead –a one-hot over $\{0, 1, \text{abstain}\}$ per member, taking the features from ten to twenty– holding the folds, the seed, the items, the estimator, the penalty, and the decision threshold fixed, so that the two arms differ in the feature map alone. Every one of the eight paired differences is unresolved. The largest in magnitude stands at 0.009 κ, under a third of the smallest effect the paper claims, and the mean at -0.001. The bound is tightest where it is most informative: on MEDHALLU, the best-powered set carrying material abstention, the interval excludes any effect beyond 0.010 κ, and on EXPERTQA, the set with the most abstentions in absolute terms, beyond 0.028. Conversely, WICE is the one set where the encoding has real room to matter and the sample no power to say so, its 8.1% abstention rate sitting on 222 items, so its interval is uninformative rather than reassuring. A missing-indicator parameterisation of the same three-level factor –the committed vote, filled to 0, beside an abstention flag– returns *identical* verdicts on every item of every dataset, so the ℓ_2 penalty's dependence on the chosen basis contributes nothing to the comparison. On this evidence, the results do not rest on the imputation, which we read as a convenience rather than as a modelling commitment. A roster abstaining far more than ours would, however, deserve the question asked again.

Multiplicity. The suite runs the same paired test over the eight datasets, so the question of what the eight verdicts mean together deserves an explicit answer. We report every interval per comparison and we read the eight as a description of where the panel stands rather than as a family of confirmatory tests. Each dataset asks its own question, and the paper's claim is the characterisation of the regimes, not a single pooled effect. The Holm check (Holm, 1979) nonetheless matters, because one verdict –the MEDHALLU win– is load-bearing, and a reader is entitled to ask whether it survives being treated as one test among several. Over the four core grounded sets, the primary denominator declared in §2,

it does: MEDHALLU's raw one-sided $p = 0.012$ adjusts to 0.048. The other three do not, and the correction bites hardest where the body already reports an unresolved difference: XSUM, the closest to resolution per comparison (one-sided $p = 0.045$), adjusts to 0.135; WICE to 0.155; RAGTRUTH to 0.280. As a consequence, the adjustment changes no verdict in the paper, since an unresolved improvement is reported as a tie in the first place. We read this as a boundary and not a reassurance. Over the widest denominator, with all eight datasets treated as one family, MEDHALLU's win would adjust to $p = 0.096$ and would not survive at the conventional level. The win is therefore a claim about the core grounded four. What the suite is evidence for is the ordering of the eight datasets, on which §3.3 rests and which no single p-value can carry.

The missing-vote denominator confound. A method's κ is scored over its own non-abstaining set: because the stacker imputes every missing member vote it is always scored on the full item set, whereas a single-judge baseline is scored only on the items where it returned a parseable verdict; whenever that baseline under-covers, the delta (stacked − best-single) therefore compares two different denominators. We surface each baseline's committed-n –i.e. the number of items it actually returned a verdict on, out of the dataset's n– beside every estimate (Table 1, †), and use the companion balanced accuracy, which charges abstention on the full denominator, as a check; where the two metrics agree, the verdict does not depend on how abstentions are handled. The one headline contrast with a materially under-covering baseline is MEDHALLU, where the CV-best judge (`Granite`) abstains on roughly 17% of items ($1{,}653$ of $2{,}000$ scored). Restricting the comparison to that common denominator leaves the gap essentially unchanged: $+0.031$ $[+0.003, +0.058]$ against the full-denominator $+0.032$ $[+0.004, +0.060]$, κ's prevalence correction absorbing the dropped items. The residual asymmetry, moreover, runs against the panel. The items `Granite` abstains on are harder and disproportionately hallucinated, so scoring the baseline only where it was willing to answer flatters it, and the reported gap understates the panel's advantage. The verdict is therefore conservative rather than inflated.

L EXTENDED RELATED WORK

LLM-as-judge and its limitations. Using a strong language model to score the outputs of another has become the default substitute for human annotation where gold labels are scarce or ill-defined (Zheng et al., 2023; Li et al., 2024a; Gu et al., 2026). Single-model judges, however, carry systematic biases –positional and verbosity effects, and a self-preference for their own generations (Panickssery et al., 2024; Chen & Goldfarb-Tarrant, 2025)– and are reported to reach human-level agreement only well above the 4–9B scale, no judge in that class clearing the bar (Han et al., 2025), which motivates cheaper judges and combinations of them.

Panels of judges. The closest line of work likewise replaces the single judge with a panel. However, it typically does so whilst asking whether the substitution pays on average, rather than under what conditions it pays. In particular, the Panel of LLM Evaluators (PoLL) reports that an ensemble of smaller judges from disjoint families can correlate with humans better than a single frontier judge, at a fraction of the cost (Verga et al., 2024). Related frameworks extend the idea with richer aggregation and to more subjective tasks (Rahmani et al., 2025; Zhao et al., 2025), whereas cascaded selective evaluation escalates to an expensive judge only where the cheap ones are unsure (Jung et al., 2025). The self-consistency family, by contrast, asks an orthogonal question: it resamples one model rather than combining judges (Manakul et al., 2023; Farquhar et al., 2024), and so catches arbitrary errors rather than the systematic ones a panel shares. Along this line, Ricco et al. (2026) find the self-consistency signal to be geometric for small models, with genuine answers to a prompt clustering more tightly in sentence-embedding space than hallucinated ones; the authors propagate labels from some thirty to fifty responses per model, where responses are judged by a frontier model — the kind of judging a cheap panel could take over.

Noisy-voter aggregation. Combining imperfect voters is a classical problem well beyond LLMs. Latent-variable models estimate each annotator's reliability without gold labels (Dawid & Skene, 1979), their descendants adding item difficulty (Whitehill et al., 2009), annotator spamming (Hovy et al., 2013), or a Bayesian prior over each annotator's confusion matrix (Kim & Ghahramani, 2012), to cite a few. Similarly, weak-supervision frameworks estimate the accuracies and correlations of many noisy sources (Ratner et al., 2017). Finally, in ensemble learning, the *diversity* of the members'

errors is the standard explanation of what combining them gains (Dietterich, 2000), although diversity measures predict that gain only weakly (Kuncheva & Whitaker, 2003). In their unsupervised use, however, all these models infer a member's reliability from its agreement with a latent consensus. On a correlated panel that consensus *is* the shared error, so the obstruction Appendix C documents for Dawid–Skene applies to them by construction, as it does to label-free aggregators designed for LLM judges (Sun et al., 2025).

Extensions that account for the dependence between members do relieve the obstruction, but only under assumptions of their own, which Appendix C spells out (Jaffe et al., 2016; Balasubramanian et al., 2026). We therefore run Dawid–Skene as the representative of the family, and leave open whether a dependence-aware extension recovers member competence on panels as correlated as ours. Furthermore, the aggregators designed for LLM judges do not escape the best single member either: on TRUTHFULQA, Sun et al. (2025) report every aggregator they compare (over 7–8B judges), their own included, below the strongest member (68.77 against 69.84). They also select their checkpoints on 250 labelled items, well above the fifty records with which our stacker already clears the vote on seven of eight sets (Appendix C).

Correlated errors. Recent work consistently reports that language models, judges among them, violate the independence assumption underlying naive ensembling. In particular, CARE (Zhao et al., 2026) traces the violation to latent confounders that the judges share and that correlate their errors. Beyond judges, Kim et al. (2025) measure the phenomenon across more than 350 models: errors overlap heavily and, decisively for a panel, correlation *rises* with model accuracy, even across providers and architectures. Goel et al. (2025) observe the same trend under a chance-adjusted error-similarity metric across 130 leaderboard models. This is what makes a panel of *cheap* judges a more interesting instrument than a panel of strong ones, and not simply a thriftier one. Indeed, the decorrelation on which aggregation feeds is most available where individual competence is lowest, which is also why our roster is drawn from disjoint families at the 4–9B scale. However, since open-ended generations are markedly homogeneous across them, disjoint families offer no guarantee of diversity *a priori* (Jiang et al., 2025). For this reason, the admissibility test reads the correlation coordinate off the calibration set instead of presuming it from the roster.

Recent panels and ensembles of LLM judges. Several recent and concurrent studies address questions close to ours. In what follows, we set each against our account, noting where it agrees and where it departs in task, judges, or aggregator. Kohli (2026) reports that a panel of nine *frontier* judges carries only about two effective independent votes on NLI tasks, falling 8–22 points below what independent voting would give, and leaving the best single judge unbeaten. We do not read this as a contrary finding. Indeed, their nine judges are worth about two independent votes, and on the four of our sets whose members are worth fewer (Appendix B) we too find that aggregation recovers the leader and no more. This is also how our account reconciles that result with the optimistic panel literature (Verga et al., 2024). Zhu & Rao (2026) likewise treat a labelling budget as the binding constraint, but for a different decision: which aggregator family, and how many judges, the available labels can support once a panel is to be run. On preference and summarisation benchmarks, with six 7–12B open judges and one API model, they report scalar or reliability aggregation beating a joint-table calibrator in 16 of 20 dataset–budget cells, current judge outputs being often "additive or redundant". This supports the redundancy premise we build on, and is why we do not pursue a richer aggregator. In the same vein, Li (2026) reports, on pairwise reward benchmarks, that a calibrated full panel beats discarding judges by accuracy, which is what our fitted coefficients show when the weakest of the ten members still draws a large weight (Appendix I).

Closer to our setting, Laddha et al. (2026) evaluate sixteen small judges (0.6–14B) for the same reasons as ours (cost, privacy, and silently changing model versions), and report that the best of ten three-judge juries drawn from their top five judges gains only 0.06 accuracy points over the best individual judge. Their leading judges share most of their mistakes, which would place those juries on the correlated side of our plane (Figure 3(a)), where the admissibility test rejects a panel. They leave decorrelated panels aside as needing "per-instance error knowledge unavailable at deployment time", whereas the test reads both coordinates off fifty to a hundred labelled records (Appendix C). Moreover, they do not price the substitution against a proprietary judge.

From a more analytical angle, Ali (2026) proposes a law for a weighted vote of two models, in which the lift over the stronger one grows with the weaker one's accuracy and with their accuracy-adjusted

complementarity, and shrinks with their accuracy gap. It is a pairwise analogue of our account, in which the gap plays the part of our dominant leader. Its claim that raw correctness correlation alone predicts almost nothing ($R^2 \leq 0.09$) is indeed consistent with our correlation coordinate being the weaker of the two when read alone (Appendix B). In a similar spirit, Kim (2026) reports that five common diversity measures largely re-express capability as predictors of majority-vote gain, leaving only a modest residual in which more shared error means less gain; this entanglement is why we report competence and error-correlation as separate coordinates.

Zhou & Lin (2026) propose the nearest analogue of our admissibility test. They flag a three-judge factuality panel as unsafe when a labelled calibration probe shows correlated false negatives, at thresholds fixed in advance, and route the panel's accepts to grounded verification. Their flag, however, reads a correlation coordinate alone, and escalates to a reference rather than to a frontier judge. Outside faithfulness, Grayzel (2026) reports three larger open-weight judges to be statistically indistinguishable from two frontier judges on pass/fail proof grading, at one to two orders of magnitude lower cost. That substitution is priced per call, however, with no calibration outlay and no account of when it holds.

Finally, and concurrently with this work, Hossain et al. (2026) report the design-effect n_{eff} of Kohli (2026) for a panel of ten judges on preference data. Those ten judges, however, are only five open models under 10B, four of them checkpoints of our roster, each run under two prompts. Among these, they report an average error correlation of 0.21, worth about 3.5 independent judges, and a stronger dependence still among three frontier judges. Their two dependence-aware filters then rank items for retention rather than weight members, and neither is reported to improve on the plain majority on eight LLM-AGGREFACT components. Conversely, we pair the correlation coordinate with a competence one, fit the weights on labels, and price the substitution of a frontier judge.

Where this work sits. Read together, these works measure how dependent a panel's errors are, choose an aggregator once a panel is to be run, or flag when a cheap panel should escalate. Our questions sit one step earlier: whether to run a cheap local panel *at all* in place of a paid frontier judge, at what price once its calibration labels are counted, and how a deployer is to tell in advance. As to adoption, the panel falls materially behind the frontier judge on only two of our eight benchmarks, both in the dominant-leader regime (§4). As for the price, the panel holds a median 93% of the frontier judge's κ at a sixty-fourth of its inference cost, and its one-time calibration breaks even between some $40{,}000$ and $81{,}000$ items. Finally, whether a new task is one where the panel can stand in is decided by two member-level conditions: the distribution of competence and the correlation of errors, both of which we test at the item level (Appendix B). The admissibility test reads both off the same calibration set (Appendix C), so a deployer can estimate before committing whether the panel can stand in. The frontier judge then needs to be kept only in the dominant-leader regime, for which the test rejects the panel. To the best of our knowledge, no work in this line gives that price or a measured account of its conditions. By contrast, we establish both under a single protocol, applied unchanged across eight benchmarks. Every difference is paired and cluster-bootstrapped, the admissibility test is checked on $1{,}888$ (sub-)panels drawn from the same members, and our harness is calibrated against the published rows it shares (Appendices C and F).

Cascading and routing. A parallel line reduces the frontier bill without removing the frontier. Each item goes to a cheap model first, and escalates only when a confidence signal says so. Chen et al. (2024) report matching a frontier model at a fraction of its cost by cascading, Jung et al. (2025) apply the idea to evaluation specifically, and it is now surveyed as a field of its own (Moslem & Kelleher, 2026). That route optimises, however, a different objective from ours, and the difference is not one of degree. Indeed, a cascade still issues frontier calls on the items it finds hard, so the evaluated text still leaves the premises, the rate limit still binds, and a retired or silently retrained endpoint still invalidates the pipeline (Hartmann et al., 2026). These are the constraints that motivate an on-premises panel in the first place, and no reduction in call volume relieves them. Within a panel, Zhu et al. (2026) learn from a small labelled audit which judges of a panel to call, on which items, and when to stop, a routing counterpart to our call-every-member stacker, which their own comparison claims to be a strong endpoint wherever its cost is acceptable. Zhang et al. (2026) route instead within one judge, between the reasoning and non-reasoning modes of a small open model under a token budget, learning the router from some $40{,}000$ labelled preference pairs — what is being routed there is compute, not the choice of judge.

Cascading and aggregation are moreover complementary rather than rival: a panel with a calibrated abstention would be a natural first stage, which is the direction our confidence-interface limitation points to (Appendix M). We therefore do not position the panel as a cheaper cascade, and where a deployer's only constraint is cost, a cascade may well be the better instrument.

Closest to our own accounting is Salinas et al. (2025), who treat cost as an explicit objective and search judge design decisions multi-objectively against it, likewise favouring open-weight judges for accessibility; however, they tune the configuration of a *single* judge for pairwise model ranking, whereas we hold a panel fixed and price the substitution it is meant to make, including the labels its aggregator consumes.

Purpose-trained judges and specialised detectors. A competing route trains a single small model expressly for evaluation. Prometheus does so for rubric-conditioned assessment (Kim et al., 2024a;b), and JudgeLM, PandaLM, and Auto-J for scalable pairwise or single-response scoring (Zhu et al., 2025; Wang et al., 2024; Li et al., 2024b). For grounded factuality, the same route builds compact verifiers: these include NLI- and question-answering-based consistency metrics (Laban et al., 2022; Zha et al., 2023), compact fine-tuned detectors such as HHEM, whose open release was trained on the RAGTRUTH training split among other data (Tamber et al., 2025), and LLM-distilled specialist checkers (Tang et al., 2024; Lei et al., 2025). The last of these are evaluated on the LLM-AGGREFACT benchmark, from whose test split four of our grounded sets are drawn (§2).

Overall, such approaches spend their labels on fine-tuning one specialist where ours go to a light second-stage aggregator over frozen off-the-shelf generalists, with no bespoke model to train or serve. They are the sharpest comparator a cheap panel has, and we treat them as one. On our class-balanced sets, a published balanced accuracy converts exactly to our primary metric as $\kappa = 2\,\mathrm{BACC} - 1$. We therefore place those anchors, specialist, guard-tuned, and frontier-prompting alike (Song et al., 2024; Kovács & Recski, 2025), beside our own panel on one scale in §4 (Table 8). The guard-tuned row is read off the leaderboard, for the reason Appendix G gives. The juxtaposition, however, carries the usual caveats, since factuality-metric comparisons are fragile to exactly these differences of protocol (Godbole & Jia, 2025). Nor do we claim that the two are interchangeable. A specialist is trained for reference-conditioned verification, sometimes on the target corpus's own training split (Tamber et al., 2025), and so holds an advantage on its home ground; yet it cannot score the ungrounded half of our suite at all. Fine-tuned judges are moreover reported to behave as task-specific classifiers that generalise poorly off their training distribution (Huang et al., 2025).

Benchmarks. We close this section with the benchmarks on which all these approaches are compared. Hallucination resists a single formal definition (Cossio, 2025; Bang et al., 2025), and the eight benchmarks we adopt accordingly span the range, from grounded faithfulness against an explicit reference to ungrounded factuality against world knowledge (§2). State-of-the-art detectors nonetheless remain far from saturating such benchmarks (Bao et al., 2025; Pandit et al., 2025), which keeps the cheap-panel question open.

M EXTENDED LIMITATIONS

The conclusion names the main limitation, i.e. the roster the panel is built from; this appendix reports the other conditions we have identified that bound –without overturning– the claims made in the main text.

Costing. Our costing is an estimate rather than a quotation. It prices GPU time at a public reference rate for the card the panel was benchmarked on, and the frontier at list API pricing on one date, so the break-even volume moves with both. It also prices metered use rather than the hardware footprint a deployer must actually provision (Appendix H). We price a label at \$1, two minutes at \$30 per hour; a domain requiring an expert annotator, of which MEDHALLU's clinical abstracts are the suite's clearest case, scales that price and the break-even volume with it. The outlay nonetheless stays one-time, so it moves where the crossing point falls and not whether there is one. The costing likewise prices inference and not the orchestration a ten-member panel needs, which a deployer serving the members sequentially pays in latency rather than in hardware (Appendix H).

Frame sensitivity and quantisation. The frame is fixed per dataset rather than tuned per judge. On the four claim-verification sets it is the better of the two frames we ran, for the frontier judge as well as for the panel (Table 11). A frontier-specific prompt search could nonetheless move the EXPERTQA reversal, which rests on the suite's weakest frontier score; few-shot exemplars, in particular, are known to lift frontier fact verifiers (Seo et al., 2025), and handing the frontier judge our calibration records in that form is a matched-label comparison we do not run. The open-judge scaling check uses 4-bit AWQ open judges, so the scaling conclusion is scoped to that quantisation (Appendix H).

External comparisons. These rest on published data obtained under their authors' own protocols: four datasets share their rows with us (Table 8), and two more are matched only in metric, on a different slice and, on RAGTRUTH, at a different prevalence (Table 9); hence on MEDHALLU we compare levels rather than ranks, and on RAGTRUTH we report only the pattern within the published anchors themselves.

We re-score six published results ourselves. Two LLM-AGGREFACT leaderboard rows agree to about a point and three MEDHALLU members to within 2.6 F1, whereas one checkpoint (`Bespoke-MiniCheck-7B`) we do not reproduce (Appendix F). A cross-paper juxtaposition nonetheless remains weaker evidence than a within-suite contrast (Godbole & Jia, 2025). The effective-independence comparison of §3.3 is across studies in the same sense: $\bar{\phi}$ and n_{eff} are panel-internal and so port between settings, but the tasks, the gold, and the panel size differ, and it is a comparison of a derived statistic and not of a score (Appendix B).

Scope of the panel and of its competitors. We reduce each judge's task to a binary faithfulness verdict and read each member through that verdict alone; graded or span-level rubrics, richer aggregators than a stacked logistic regression, and interfaces that expose a member's confidence rather than its decision are all out of the current scope, and the last of these bounds the panel from a direction we do not measure. Our competitors are off-the-shelf judges, other aggregators of them, and the purpose-trained detectors, which we run ourselves across the suite (Appendix G); we do not, however, re-train a specialist on the labelling budget the stacker consumes. Although we sweep panel *size* (Figure 6(a)), we do not vary its *composition*. We thus characterise when aggregating one fixed roster of frozen generalists helps, not whether another roster, or a matched-budget specialist, would do better. Symmetrically, the frontier side of the comparison is a single judge, CLAUDE SONNET, so the substitution is priced against that model and not against frontier judges at large. Nor do we measure an escalation cascade, which keeps the frontier judge for the items a cheap model is unsure of. We read it as complementary to the panel rather than as a competitor (Appendix L), and leave open what share of frontier calls a panel-first cascade would retain at matched quality.

Class balance. Every set is built at an exact $50/50$ balance (Appendix J), which a deployment stream need not share. Under a shift in prevalence alone, each member's per-class recalls held fixed, balanced accuracy is unchanged by construction, whereas κ is not, depending on prevalence even at fixed recalls, and the identity $\kappa = 2\,\text{BACC} - 1$ no longer holds. The stacker, fitted on a balanced calibration sample and read at 0.5, targets to a first approximation the balanced rather than the plain error, so a deployer minimising the latter would shift its intercept by the log-ratio of the deployment priors or, more simply, draw the calibration sample from the deployment stream itself. We evaluate neither remedy, so our κ readings hold for balanced evaluation only, and so does the admissibility rule, whose two statistics are computed on the same balanced samples.

Evidential base. The account is supported at the item level over nearly two thousand points and within every dataset (Appendix B), but the *cross-dataset* ordering of the differences rests on eight datasets, so any dataset-level competence/correlation predictor is illustrative rather than confirmatory. The account and the admissibility rule were moreover formed on these same eight datasets, and with a single clear gain over the best single judge in the suite they are consistent with it rather than validated by it. Furthermore, the suite never leaves English, its eight sets being drawn from biomedical, encyclopaedic, news, and open-domain question-answering sources, so whether the two statistics keep their reading in other languages or domains is untested. Finally, both Cohen's κ and balanced accuracy inherit whatever label noise the underlying benchmarks carry.