DOES PROMPT SENSITIVITY IMPROVE CROSS-FAMILY DEGRADATION PREDICTION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Models whose answers barely change when instructions are reworded might also hold their accuracy when the data change. Does prompt sensitivity predict how much accuracy a new model family will lose, beyond source accuracy? We test two source-side features, accuracy dispersion and label-free disagreement across ten wordings, on fifteen open-weight models, three classification tasks and seven natural shifts. Both lower held-out error under cross-validation over five development families. Under one frozen fit, disagreement raises error by 6% in a first held-out cohort and lowers it by 35% and 30% in a second and a third, of three and two families; dispersion has the same signs. No cohort reaches the registered level, and no test pools them. Disagreement's falls in the second and third cohorts exceed a within-source permutation reference, but in development the invalid-response rate alone predicts degradation nearly as well, and the first cohort's rise reverses when one training family is dropped. The evidence does not establish that the gain transfers to a new family, nor that sensitivity carries no information about shift.

1 INTRODUCTION

A team that has measured a set of models on the shifts it cares about, and then acquires a new model, wants to know what that model will lose on those shifts without evaluating it on every one of them. What it can compute cheaply about the new model lives on the source side. One such signal already works: accuracy in distribution predicts accuracy out of distribution across large collections of models and shifts (Miller et al., 2021; Taori et al., 2020), well enough to be a baseline, although the relation bends and sometimes breaks on language tasks (Yuan et al., 2023; Teney et al., 2023; Wenzel et al., 2022). Prompt sensitivity is a candidate to add. It is computed on the source side, it tracks accuracy closely enough to have been proposed as an unsupervised proxy for performance (Lu et al., 2024), and in the form that counts how often two wordings disagree on an item it needs no labels. The hope is easy to state: a model whose answers barely move when the instructions are reworded might also hold its accuracy when the data change. We ask whether prompt sensitivity improves the prediction of degradation beyond source accuracy, for a model family the predictor has never seen.

The obvious test answers a different question. Falcon3 7B is the most accurate of our seven development models, 85.8 points, and its accuracy barely moves across ten wordings, a standard deviation of 1.6; Qwen3 4B is less accurate, 76.6, and about as stable, 1.9. Ranked by stability, Falcon3 7B looks at least as safe, yet under shift it loses 19.7 points against Qwen3 4B's 10.7. Across the seven models dispersion falls as source accuracy rises, with a rank correlation of -0.68, and a more accurate model has more accuracy to lose, so a raw association between sensitivity and degradation can be produced by accuracy alone. The comparison therefore has to be incremental over source accuracy and grouped by model. And because cross-validation over five families reuses each family in the other folds' training sets, the families that decide the question have to be held out of the fit altogether.

One result held in every held-out cohort: source accuracy with task and shift indicators predicts a new family's degradation on the shifts seen in training, with held-out R^2 of 0.76, 0.64 and 0.81 (Section 5.1). Figure 1 is the result on the question. Under cross-validation over the development families, adding either feature lowers held-out error in most families. Three further families, held out under a frozen predictor, go the other way: error rises in two of them. A second cohort of three

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

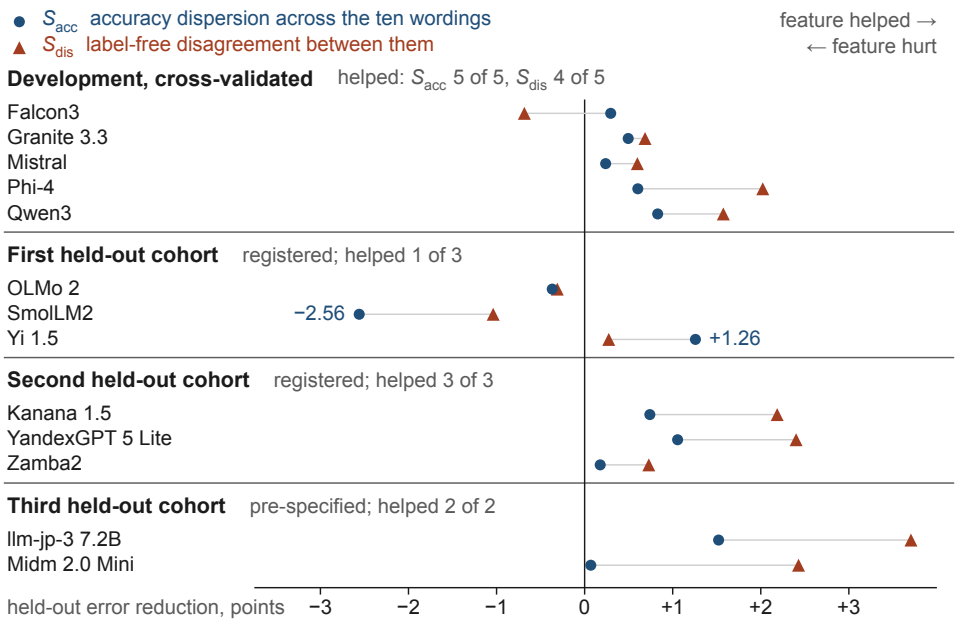


Figure 1: Per family, the fall in held-out mean absolute error, in points, when S_{acc} or S_{dis} is added to the source-accuracy predictor, positive when the feature helped. Development families are left out in turn; the three held-out cohorts are predicted by one frozen fit on all development cells and are shown on one axis, not pooled.

new families, predicted by the same frozen fit, goes back: error falls in all three. A third cohort of two families, pre-specified from the second cohort's remaining candidates, falls again. Checks run afterwards find that the first cohort's estimate moves when one training family is dropped or malformed responses are rescored, and a pre-specified within-source permutation reference contains the first cohort's rises but not disagreement's falls in the second and third cohorts (Section 6.4).

The claim this supports is bounded. A gain from prompt sensitivity seen under cross-validation, or in one held-out cohort, did not carry to all the others, so the evidence does not establish that the gain transfers to a new family. The source-accuracy baseline does predict a new family on shifts seen in training, better than task and shift indicators alone where compared (cohorts one and two); a shift absent from training has no fitted intercept under this design. We did not find the incremental comparison for an unseen family in prior work (Section 2).

Contributions.

- **An incremental, family-held-out test** of two sensitivity features over a source-accuracy baseline (Sections 3 and 4).
- **Three held-out cohorts under one frozen predictor**: error rises in the first, falls in the second and in the two-family third, none at a resolution that settles the sign (Section 5).
- **A within-source permutation reference and a residualisation check of the increment**, pre-specified (Section 6.4).
- **A measurement of how far the estimate moves** with the training pool, the scoring rule, the baseline's calibration and the wording sample, and a split of the label-free feature into its valid and malformed parts (Sections 5.6 and 6).

2 Related Work

Sensitivity as a measured property. That prompts change accuracy is established and quantified: formatting changes alone move accuracy by as much as 76 points (Sclar et al., 2024), single-template evaluation is unstable (Mizrahi et al., 2024; Voronov et al., 2024), and the variation is now a reportable index (Chatterjee et al., 2024; Errica et al., 2025; Zhuo et al., 2024) that can be estimated

from fewer evaluations than a full sweep (Polo et al., 2024). We ask a downstream question: whether the measured variation predicts an outcome under shift once source accuracy is known.

Sensitivity as a proxy for something else. Lu et al. (2024) perturb the input example, summarize how often the prediction changes with a variation ratio, and report a strong negative relation with accuracy ($r = -0.88$ with top-k sampling, -0.55 greedy), proposing the score as an unsupervised proxy for performance; Cox et al. (2025) use paraphrase sampling to improve uncertainty calibration. The first establishes what we take as given, that sensitivity tracks accuracy; neither tests whether it adds anything once source accuracy is in the predictor. Our disagreement feature perturbs the instruction wording rather than the input and counts pairwise disagreement rather than modal frequency, so our result is about ours alone.

Predicting accuracy under shift. Our baseline is the accuracy-on-the-line relation (Miller et al., 2021), supported at scale (Taori et al., 2020), dataset-dependent (Wenzel et al., 2022), and studied for language shifts where it can bend or fail (Yuan et al., 2023; Teney et al., 2023; Koh et al., 2021). Agreement-on-the-line (Baek et al., 2022) estimates shifted accuracy from the agreement of independently trained models on unlabeled target examples, and Saxena et al. (2024) carry it to fine-tuned foundation models. Both read agreement across models on the target side; ours is within one model, across wordings, on the source side, calibrated on other models' measured degradation.

Is measured sensitivity a scoring artifact? Hua et al. (2025) argue that much of what is reported as prompt sensitivity is produced by rigid scoring rather than by the model's handling of the task. An invalid response is itself a label value in our disagreement feature, so the concern applies to it most directly; Section 6.2 rescores the same responses under a more lenient rule.

3 THE TEST

3.1 FOUR QUANTITIES

A *cell* is one model evaluated on one source split and the target split it shifts to. Within a cell the model answers the same items under ten wordings of the same task (Section 4.2). *Source accuracy* A_{src} is its accuracy on the source split, averaged over the ten wordings; *accuracy dispersion* S_{acc}, a dispersion and not a source accuracy, is the standard deviation, in accuracy points, of that per-wording accuracy; *prediction disagreement* S_{dis} is computed per item as the fraction of the 45 wording pairs that extract different labels, then averaged over items, with a malformed response treated as its own label value; and *degradation* D is the mean over the ten wordings of source accuracy minus target accuracy, so a positive value is accuracy lost. Appendix A computes both features on one real cell. A_{src} and S_{acc} need source labels; S_{dis} needs none, since two wordings can be seen to disagree without knowing which is right. None needs anything from the shifted distribution, although the predictor they enter is fit on other models' measured degradation.

3.2 AN INCREMENTAL COMPARISON, GROUPED BY FAMILY

The two predictors are

$$\text{baseline: } D \sim A_{\mathrm{src}} + \text{task} + \text{shift}, \qquad \text{augmented: } D \sim A_{\mathrm{src}} + \text{task} + \text{shift} + F,$$

with F either S_{acc} or S_{dis}, task group and shift entering as indicator variables, and the baseline a pooled, raw-scale form of the accuracy-on-the-line predictor (Miller et al., 2021). Both are ridge regressions whose penalty is chosen inside each training fold by grouped cross-validation over families (Appendix B); numeric features are standardized with training-fold statistics, and an indicator level absent from a training fold is set to zero at prediction time. For a held-out cohort the registered statistic is the per-family fall in held-out mean absolute error (MAE) when F is added, tested by an exact one-sided sign-flip test over families at level 0.025; tables and text also give the pooled relative change in MAE, positive when the feature helped, and the number of families helped, as descriptions. A relative improvement is read only where the baseline's own held-out R^2 is positive. Every split is grouped by model: the source-side quantities are shared by every shift leaving a source, so a split with some cells of a model in training and others in evaluation would let the predictor recognize the model rather than use the feature.

3.3 PROTOCOL

Seven models from five families, Falcon3, Granite 3.3, Mistral, Phi-4 and Qwen3, on seven shifts, 49 cells, form the *development pool*, where the predictor is fit and its penalty tuned. Two cohorts of three further families are held out: OLMo 2, SmolLM2 and Yi 1.5 (19 cells), then Kanana 1.5, YandexGPT 5 Lite and Zamba2 (20 cells), and a pre-specified third of two, Midm 2.0 Mini and llm-jp-3 (13 cells; Section 5.5); a model loses a shift whose prompts exceed its context (Appendix B). Each cohort is predicted by one fit on all 49 development cells, with no held-out label entering any fit or tuning decision.

The analysis, from feature definitions to reading rules, was frozen on 22 September as a list of file hashes, before the development analysis was run and before any held-out output was opened. That list is kept in the project's own version control, which it entered on 23 September; no external registry holds it or the later registrations, so their dates rest on our records. Before the freeze the analysis code had been run on partial development output (three to four families, 28 cells) to debug it; no design choice was taken from that run. One analysis file was edited after the freeze without the dated amendment the protocol required: before the first cohort's analysis its family map gained the held-out model identifiers, without which the three families enter the test as one unit, and on 23 September it gained output fields for MAE. The amendment was written on 24 September, after that analysis, with a check that every frozen estimate is unchanged to within 10^{-12} (Appendix B). The second cohort was registered on 24 September, after the first cohort's result and before any of its own output was opened, with the predictor unchanged and a clause forbidding pooled tests; the third was pre-specified on 25 September from the remaining candidates of the second cohort's registered list, screened under the amended serving harness of Appendix B, with a rule that reads a rise and a fall alike.

Registration fixed the predictor, the comparator and each cohort's reading before its output was read; it did not buy resolution. Three families passed the first cohort's screen, the second cohort was registered at the same size, and with three a sign-flip test cannot go below a one-sided p of $1/8$, short of the 0.025 level, so each cohort is a description with its exact p and families helped. The second cohort's rule has no outcome for a fall in error, so a fall of at least 0.30 points returns "mixed".

Three labels mark the status of every analysis. *Registered*: fixed before the output it reads was opened, as the primary analyses above. *Pre-specified*: run after some cohort results were known, with plan, rule and threshold written first, as the controls of Section 6.4 and the third cohort. *Descriptive*: post hoc, no rule in advance. None replaces a primary result.

4 MEASUREMENT

4.1 MODELS, SHIFTS AND SCREENING

The development pool holds Qwen3 1.7B and 4B, Granite 3.3 2B, Falcon3 3B and 7B, Phi-4-mini, and Mistral 7B, two families contributing two models each so that a family holdout removes more than one. The first held-out cohort is OLMo 2 7B, SmolLM2 1.7B and Yi 1.5 9B; the second, Kanana 1.5 8B, YandexGPT 5 Lite 8B and Zamba2 7B; the third, Midm 2.0 Mini and llm-jp-3 7.2B. Across the fifteen evaluated models, source accuracy ranges from 56.8 to 85.8 points, dispersion from 1.3 to 9.2, and mean degradation from 2.6 to 21.7 (Table 5). Candidates were screened under the canonical wording only, on the first twenty items of each of the three source splits (60 requests), which are excluded from every evaluation; a candidate with a valid response to fewer than 90% of them was excluded before any evaluation item was scored. The criterion concerns the output contract only; nothing else was computed for an excluded candidate. It excluded two development candidates (Appendix C), three of six for the first cohort, and three of the nine screened for the second, and two of the four screened for the third, four others being gated and unfetched (Section 5.5).

Seven shifts span three task groups, all from an existing benchmark suite of distribution shifts (Yuan et al., 2023): sentiment on Amazon reviews shifting to DynaSent, SemEval and SST-5; toxicity on Civil Comments shifting to its adversarial split; and natural language inference on MNLI shifting to ANLI, ContractNLI and WANLI. Each split holds 200 items; the first twenty are the screening items, so evaluation uses 180 per split and wording. One further inference shift, MNLI to SNLI, was held back as a descriptive check (Appendix D).

Table 1: Held-out mean absolute error (MAE) of predicted degradation, in accuracy points, without and with a sensitivity feature. Change: relative fall in MAE when the feature is added, and the units (families or shifts) in which it fell, computed from unrounded MAE. The three held-out cohorts are predicted by one fit on the 49 development cells and are not pooled; the third cohort (two families) is the second cohort's remaining list screened under the amended harness (Appendix E).

		Baseline		$+S_{\mathrm{acc}}$		$+S_{\mathrm{dis}}$	
Held out	Cells	MAE	R^2	MAE	Change	MAE	Change
A new family, on shifts seen in training							
Development, one family out	49	6.15	0.41	5.64	$+8.3\%$, 5/5	5.42	$+11.8\%$, 4/5
First cohort, registered	19	6.58	0.76	7.24	-10.1%, 1/3	6.97	-6.0%, 1/3
OLMo 2 7B	6	5.38		5.75		5.69	
SmolLM2 1.7B	7	8.75		11.31		9.79	
Yi 1.5 9B	6	5.25		3.99		4.97	
Second cohort, registered	20	5.21	0.64	4.52	$+13.1\%$, 3/3	3.38	$+35.0\%$, 3/3
Kanana 1.5 8B	7	4.49		3.75		2.30	
YandexGPT 5 Lite 8B	7	6.30		5.25		3.90	
Zamba2 7B	6	4.77		4.59		4.04	
Third cohort, pre-specified	13	10.07	0.81	9.33	$+7.3\%$, 2/2	7.05	$+30.0\%$, 2/2
llm-jp-3 7.2B	6	12.47		10.95		8.76	
Midm 2.0 Mini	7	8.01		7.94		5.59	
A new shift							
Development, one shift out	49	9.10	0.14	9.51	-4.5%, 2/7	9.66	-6.2%, 2/7

4.2 Ten wordings of one task

Each task has ten wordings: one canonical template and nine built by stacking one to three changes drawn from six surface axes, namely the mood of the instruction, lexical choice, the order in which labels are listed, whitespace, the delimiter between fields, and the order of the sections. Every wording keeps the task definition, the label set and its meanings, the role of the input, and the response contract. Before use, two reviewer models from different families, Qwen3 8B and Llama 3.1 8B, judged each template while masked to the hypothesis, to which template was canonical and to every result, alongside a structural check and a sentence-embedding cosine floor of 0.85; all thirty templates passed, the non-canonical ones at cosine 0.9218 to 1.0000 from their canonical counterpart. This is a screen against large drift in meaning, not a proof of equivalence.

4.3 The response contract

A request asks for valid JSON carrying the item identifier and label. A response that does not conform scores its item as incorrect, and no response is retried. Every request carries a single item, decoding is greedy, and weights are served at half precision (Appendix B).

Accuracy under a response contract mixes task correctness with format compliance, so both features contain some format behavior (Hua et al., 2025). Among the thirteen models of the development pool and the first two cohorts, invalid responses stay at or below 0.5% on both sides for seven and are highest for Falcon3 3B, 5.2% on the source side and 10.0% on the target side, which also has the highest dispersion; the third cohort's llm-jp-3 is higher on the source side (Table 5). Over those thirteen models the source-side invalid rate correlates with dispersion at $r = +0.74$ and with disagreement at $r = +0.55$.

5 What the held-out families returned

5.1 Source accuracy predicts a new family on seen shifts

For a new family on the seven shifts the predictor was trained on, the baseline works (Table 1): its held-out error is 6.58 points on the first cohort and 5.21 on the second, against 13.02 and 8.62 for predicting the training mean, with held-out R^2 of 0.76 and 0.64, and 10.07 on the two-family third cohort (R^2 of 0.81; Section 5.5). A pre-specified comparator with the task and shift indicators alone reaches 7.40, 11.12 and 5.35 points in development cross-validation and on the first two cohorts (it

was specified before the third and was not computed on it), so adding source accuracy lowers error in all three evaluations (Table 2): by 1.25 points in development, by 4.54 points on the first cohort, most of it on SmolLM2 (17.93 against 8.75), and by 0.14 on the second. Per family, the indicators alone do better in two of eleven cases, Falcon3 in development (5.74 against 7.06) and YandexGPT in the second cohort (5.17 against 6.30). By construction the baseline has no intercept for a shift absent from training, and on the one shift held back from every fit it does worse than the training mean (Appendix D), so the seen-shift result is a within-support statement.

5.2 Under cross-validation, the features lower error for a withheld family

Withhold one development family at a time and the hope appears to hold. Adding accuracy dispersion lowers held-out error for all five families, 6.15 to 5.64 points ($+8.3\%$); adding prediction disagreement lowers it for four of five, to 5.42 ($+11.8\%$; Figure 1, top). The folds share training families, so the five are not independent tests. Withholding one shift at a time instead, both features raise error (Table 1).

5.3 The first held-out cohort: error rises

On the first cohort, 19 cells, adding accuracy dispersion moves held-out error from 6.58 to 7.24 points (-10.1%) and adding disagreement moves it to 6.97 (-6.0%), each helping one family of three (one-sided $p = 0.75$ and 0.875, resolution-limited), rises within the within-source permutation reference of Section 6.4 (placebo mean -0.15 and -0.19, $p_{\mathrm{harm}} = 0.149$ and 0.305, $N = 4000$; pre-specified); disagreement's direction holds in only 63 of 84 wording subsets and depends on the wording sample. The rise is concentrated in SmolLM2, which goes from 8.75 to 11.31 points with dispersion; its source accuracy (57.0 points) and degradation (2.6) both lie outside the development ranges (73.4 to 85.8, 10.7 to 21.7), which contain all three second-cohort models; it lies below them by 16.4 and 8.1 points. OLMo 2 moves by less than half a point, and Yi 1.5, whose degradation (9.6) is also below that range, improves. In the second cohort, YandexGPT's dispersion and disagreement sit just below the development ranges (Table 5).

5.4 The second held-out cohort: error falls in every family

Under the same frozen fit the first cohort's direction does not repeat. Adding dispersion moves error from 5.21 to 4.52 points and disagreement to 3.38, each lowering it in all three families ($p = 0.125$), resolution-limited. Disagreement's fall exceeds its within-source permutation reference, larger than every one of the 4000 permuted columns (placebo mean -0.07, $p = 0.0002$), and holds in all 84 wording subsets; dispersion's fall lies within its reference (placebo mean -0.11, $p = 0.078$; pre-specified, Section 6.4). The rule registered for this cohort was written to ask whether the first cohort's harm repeats, so it has an outcome for that and none that says the feature helps; it labels any fall of at least 0.30 points "held-out evidence is mixed" (Appendix E). Neither cohort alone, nor the two together, settles the sign of the benefit.

5.5 A third cohort, pre-specified

The third cohort was pre-specified on 25 September from the remaining candidates of the second cohort's registered list, screened under the amended serving harness of Appendix B: of 8 candidates in registered order, 2 passed, Midm 2.0 Mini and llm-jp-3 7.2B, 2 failed the format criterion and 4 are gated repositories that could not be fetched (Appendix E), so the cohort has two families and 13 cells. Under the same frozen fit the baseline reaches 10.07 points (R^2 of 0.81); dispersion lowers error to 9.33 ($+7.3\%$) and disagreement to 7.05 ($+30.0\%$), in both families (Table 1, Figure 1). The rule, written after both earlier cohorts' directions were known, reads a rise and a fall alike; at two families it returns a description, not a test (one-sided p floor 0.25, resolution-limited), unpooled with the earlier cohorts. The wording resampling and the checks of Sections 6.1 to 6.3 predate this cohort and are not computed on it; its controls are the permutation reference and the residualisation check. Of three held-out cohorts predicted by one frozen fit, disagreement lowered error in the second and the third and raised it in the first; none can reach the registered level, so the direction is favourable in two of three descriptions and established in none. Disagreement's fall exceeds its within-source permutation reference (placebo mean -0.13, $p = 0.0020$, $N = 4000$) and dispersion's lies within

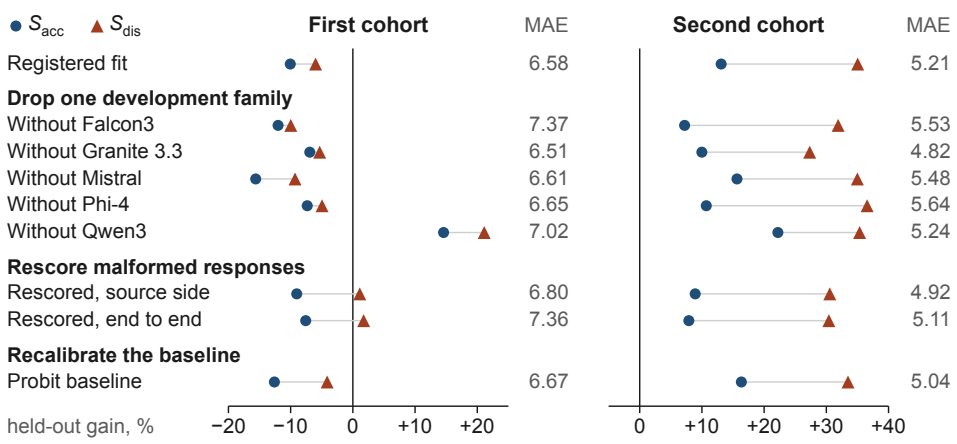


Figure 2: Held-out gain from adding S_{acc} (circles) or S_{dis} (triangles), in percent, positive when error fell, relative to that row's own baseline, whose MAE is printed at the right of each panel. Rows change one thing: the development families in the fit, the rule scoring malformed responses, or the scale of the baseline. The spread is sensitivity to the procedure, not a confidence interval.

it ($p = 0.062$); the invalid rate is nonzero in 6 of 13 cells, which limits what the residualisation check can show (Section 6.4). Both families lie outside the development range of source accuracy (16.6 and 4.4 points below), as SmolLM2 does, yet both features lowered their error, so position outside the range does not by itself explain the first cohort's rises; over the eight held-out families the rank correlation of the per-family change with that distance is -0.03 for dispersion and $+0.30$ for disagreement (a description of eight points, as are the eight-family intervals of Section 6.4).

5.6 Where the label-free feature's development gain sits

Disagreement needs no labels, but it counts a malformed response as a label of its own. A pre-specified split divides it exactly into wording pairs that give two different valid labels and pairs where one response is invalid, each entered alone in place of S_{dis} (Appendix F). Its pre-specified reading is "mixed", because the invalid part is on average only 15% of the statistic. Descriptively, that part alone recovers 79% of the gain and the valid part none: most of the development gain is recovered by the part that counts malformed responses. The pattern does not repeat on the second cohort, where the two parts recover 34% and 40% of its gain, and neither part lowers the first cohort's error. The parts are fit separately, so the shares need not add up; in 13 of the second cohort's 20 cells the invalid part is zero, so its predictions there change only through coefficients refit on development. A pre-specified residualisation check agrees for development (Section 6.4).

6 What moves the estimate

The second cohort's gains remain positive under training-family deletion, rescoring and probit calibration; their magnitudes move. Figure 2 changes one thing at a time: the development families in the fit, the scoring of a malformed response, and the scale on which source accuracy enters the baseline; all three are pre-specified, the first two reaching the second cohort only after their first-cohort results were known.

6.1 The training pool

On the first cohort four of the five removals leave both changes negative; removing the two-model Qwen3 family reverses both, to $+14.6\%$ for dispersion and $+21.1\%$ for disagreement. Removing any single development model leaves both negative, and over the 21 pairs of development models (Appendix G) the four pairs that reverse the label-free sign all contain Qwen3 1.7B, and none of the ten pairs without a Qwen3 model does. On the second cohort no removal changes a sign: disagreement's gain stays between $+27\%$ and $+37\%$ and dispersion's between $+7\%$ and $+22\%$, with error falling in all three families every time.

Table 2: Held-out MAE in points for the registered predictors, two simpler comparators and a probit calibration of source accuracy (pre-specified). Development: one family out; cohorts: one fit on the 49 development cells. The comparators and the calibration were specified before the third cohort and were not computed on it.

Predictor	Development	First cohort	Second cohort
Training mean	9.42	13.02	8.62
Task and shift indicators only	7.40	11.12	5.35
Registered baseline: A_{src} + indicators	6.15	6.58	5.21
$+S_{\mathrm{acc}}$	5.64	7.24	4.52
$+S_{\mathrm{dis}}$	5.42	6.97	3.38
Probit baseline	6.67	6.67	5.04
$+S_{\mathrm{acc}}$	5.83	7.52	4.22
$+S_{\mathrm{dis}}$	5.60	6.95	3.35

6.2 The scoring rule

Rescoring malformed responses reverses disagreement's gain in development and moves the first cohort's across zero; in the second cohort it rescues few responses. The alternative rule recovers a label only from a rejected response that names exactly one allowed label unambiguously (Appendix H). Over the development pool and the first cohort it recovers 1,790 of 5,225 rejected responses. In development disagreement's gain goes from $+11.8\%$ to -12.9% when features and outcome are both rescored, while dispersion's shrinks from $+8.3\%$ to $+3.8\%$. On the first cohort, rescoring only the source side, with the outcome held at the registered degradation, moves disagreement's change across zero, to a gain of 0.075 points computed from unrounded values (6.80 to 6.73); dispersion stays negative. Both features still lower error in all three second-cohort families under either rescoring, from 4.92 to 4.48 with dispersion and 3.42 with disagreement when the source side is rescored, and from 5.11 to 4.71 and 3.56 when the outcome is rescored too (Table 9; rescue counts per model in Appendix H).

6.3 The calibration of the baseline

Accuracy on the line is usually drawn on a probit scale (Miller et al., 2021), which extrapolates differently from the registered linear baseline. A pre-specified alternative, specified before the third cohort and not computed on it, regresses probit target accuracy on probit source accuracy and the same indicators, with the same estimator, and maps predictions back to points of degradation (Appendix I). It makes the baseline worse in development, 6.67 against 6.15 points, leaves the first cohort almost unchanged, 6.67 against 6.58, and improves the second, 5.04 against 5.21 (Table 2); on SmolLM2 it is worse, 9.29 against 8.75. Added to it, disagreement raises first-cohort error by 0.28 points and lowers second-cohort error by 1.69. The pre-specified criterion, at least 0.25 points in each cohort, calls the opposite directions persistent, on the first cohort by a margin of 0.03 points; the first cohort's rise is smaller than under the registered baseline, 0.28 points against 0.39.

6.4 Two controls

Two pre-specified controls give the increment a reference (Table 3; details and the six- and eight-family intervals in Appendix J). The first permutes each feature across models within block (development and held-out together) and source dataset, 4000 times, and repeats the frozen fit: the reference keeps the feature's per-source marginals and breaks its association with the model, hence with source accuracy and invalid rate, so it is not a test of information beyond accuracy; it is read in both directions at 0.05 without multiplicity correction. Both development gains exceed it ($p = 0.0027$ for dispersion, 0.0032 for disagreement); the first cohort's increases lie within their references ($p_{\mathrm{harm}} = 0.149$ and 0.305); in the second cohort disagreement's fall exceeds its reference, larger than every one of the 4000 permuted columns ($p = 0.0002$), while dispersion's lies within it ($p = 0.078$), and the third cohort reads the same way ($p = 0.0020$ and 0.062; Appendix J). The second enters the source-side invalid-response rate alone, and each feature after removing its training-fitted linear association with that rate. In development the invalid rate alone lowers error by 0.71 points, against 0.73 for disagreement and 0.51 for dispersion; dispersion's residual keeps 8%

Table 3: Two pre-specified controls, in points of held-out MAE, positive when error fell. Placebo: the feature's gain beside the mean gain of 4000 columns permuting it across models within block and source dataset, and the one-sided placebo p in the direction of the observed change (seeds in Appendix J). Compliance: the gain from the source-side invalid-response rate alone (INV) and from each feature residualised on it (Res., with the share of the feature's gain kept); Cells: cells with a nonzero invalid rate. The held-out Res. values are printed without a verdict: the first cohort has no gain to attribute (its raw changes are negative) and the second and third are degenerate (†: fewer than half the cells nonzero), so none of them is evidence about what a feature carries.

	Placebo, $+S_{\mathrm{acc}}$			Placebo, $+S_{\mathrm{dis}}$			Compliance			
Block	Gain	Mean	p	Gain	Mean	p	INV	Res. S_{acc}	Res. S_{dis}	Cells
Development	$+0.51$	-0.64	0.0027	$+0.73$	-0.47	0.0032	$+0.71$	$+0.04$ (8%)	$+0.51$ (70%)	32/49
First cohort	-0.66	-0.15	0.149	-0.39	-0.19	0.305	-0.46	$+0.43$	-1.00	13/19
Second cohort	$+0.68$	-0.11	0.078	$+1.82$	-0.07	0.0002	$+0.78$	$+0.02$	-0.47	7/20†
Third cohort	$+0.74$	-0.14	0.062	$+3.02$	-0.13	0.0020	$+2.28$	$+0.96$	-9.56	6/13†

of its gain, disagreement's 70%. Held out, the residualised disagreement gives -1.00, -0.47 and -9.56 points, printed without a verdict: the first cohort had no gain to attribute (raw changes below the 0.15-point rule threshold), the second and third are degenerate (invalid rate nonzero in 7 of 20 and 6 of 13 cells), and none is evidence about what disagreement carries. In every block the invalid rate alone moves error in the direction of disagreement's change ($+0.71$ against $+0.73$ in development, -0.46 against -0.39 in the first cohort, $+0.78$ against $+1.82$ in the second, $+2.28$ against $+3.02$ in the third), so the pattern across cohorts is one that a single format scalar also reproduces; disagreement's gain is the larger in the second and third. In development and the first two cohorts, resampling the ten wordings (84 subsets) keeps each block's direction in at least 76 of 84 except dispersion in development (66) and disagreement in the first cohort (63); bands in Appendix J.

What the checks show. Neither training-family deletion nor probit calibration reverses the second cohort's gains; the first cohort's estimate moves with the training pool and the scoring rule. The first cohort's rise is concentrated in SmolLM2, the one model of the first two cohorts outside the development range of source accuracy: with its cells set aside, dispersion's change on that cohort is positive and disagreement's near zero (Appendix J). This locates the rise without explaining it. Of the held-out changes, only disagreement's falls in the second and third cohorts exceed their permutation references, and in development the invalid-response rate alone predicts nearly as well as disagreement.

7 Conclusion

Prompt sensitivity looked like a useful addition to source accuracy when development families were held out in turn, by more than a within-source permutation of the feature gives. With families held out prospectively under one frozen predictor, it raised error on one cohort of three families, within that permutation reference, and lowered it on the next two, of three and of two families, for disagreement beyond the reference in both and for dispersion within it; the two third-cohort models lie outside the development range of source accuracy, so the direction there is not evidence of the predictor working inside its support; in development the source-side invalid-response rate alone predicts degradation nearly as well as disagreement. The gain from sensitivity is therefore an estimate tied to the evaluation that produced it: this evidence does not establish that it transfers to a new family, and it does not show that sensitivity carries no information about shift. Source accuracy with indicators predicts a new family's degradation on shifts seen in training; a new shift has no fitted intercept under this design.

Three things would settle more: one held-out cohort of at least six families (Appendix B); more development families, so that no one family moves the estimate; and a view of where each new family sits in the training support, task by task.

REFERENCES

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J. Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. POSIX: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14550–14565, Miami, Florida, USA, 2024. Association for Computational Linguistics.

Kyle Cox, Jiawei Xu, Yikun Han, Rong Xu, Tianhao Li, Chi-Yang Hsu, Tianlong Chen, Walter Gerych, and Ying Ding. Mapping from meaning: Addressing the miscalibration of prompt-sensitive language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23696–23703, 2025.

Federico Errica, Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco. What Did I Do Wrong? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1543–1558, Albuquerque, New Mexico, 2025. Association for Computational Linguistics.

Andong Hua, Kenan Tang, Chenhe Gu, Jindong Gu, Eric Wong, and Yao Qin. Flaw or artifact? Rethinking prompt sensitivity in evaluating LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19889–19899, Suzhou, China, 2025. Association for Computational Linguistics.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5637–5664, 2021.

Sheng Lu, Hendrik Schuff, and Iryna Gurevych. How are prompts different in terms of sensitivity? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5833–5856. Association for Computational Linguistics, 2024.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 7721–7735, 2021.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of What Art? A Call for Multi-Prompt LLM Evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. doi: 10.1162/tacl_a_00681.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. Efficient multi-prompt evaluation of LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Rahul Saxena, Taeyoun Kim, Aman Mehra, Christina Baek, Zico Kolter, and Aditi Raghunathan. Predicting the performance of foundation models via agreement-on-the-line. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Damien Teney, Yong Lin, Seong Joon Oh, and Ehsan Abbasnejad. ID and OOD performance are sometimes inversely correlated on real-world datasets. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e304d374c85e385eb217ed4a025b6b63-Abstract-Conference.html.

Anton Voronov, Lena Wolf, and Max Ryabinin. Mind your format: Towards consistent evaluation of in-context learning improvements. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6287–6310, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Florian Wenzel, Andrea Dittadi, Peter V. Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, Bernhard Schölkopf, and Francesco Locatello. Assaying out-of-distribution generalization in transfer learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/2f5acc925919209370a3af4eac5cad4a-Abstract-Conference.html.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, FangYuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in NLP: Benchmarks, analysis, and LLMs evaluations. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, 2023.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1950–1976, Miami, Florida, USA, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.108/.

STATEMENT ON THE USE OF LANGUAGE MODELS

The research ideas originated with the authors. Language-model assistants helped draft and edit the manuscript, improve grammar and clarity, develop analysis code, and check results and references. Language models also performed the prompt-equivalence and extraction audits described in Section 4.2 and Appendix H; these are automated assessments, not human validation. The authors retain responsibility for the methods, results, and conclusions.

REPRODUCIBILITY STATEMENT

The supplementary material provides analysis code, model-level cell tables, prompt templates, split manifests, and protocol documents. Appendix B and Tables 4–5 specify the model revisions, inference settings, data selection, and scoring rules. These materials support reproducing the reported analyses and rerunning the model evaluations.

A THE TWO FEATURES ON ONE CELL

Figure 3 computes both features of Section 3.1 on one development cell, from the recorded responses.

B THE FROZEN SETUP

The ridge penalty is chosen from $\{0.01, 0.1, 1, 10, 100\}$ by grouped cross-validation over the training families nested inside each fold; nothing else is tuned. Of the frozen files, one no longer matches its hash, the analysis module, for the two edits of Section 3.3, the first at a time not recorded before the first cohort's registered analysis; the protocol required a dated amendment before any held-out output was opened, and the analysis-module amendment of 24 September carries a runnable check that loads the frozen and the edited module and finds every frozen output identical to within 10^{-12}.

The harness amendment, dated 25 September and written before any model was screened or run under the amended harness, changes one line of the serving harness: the token-type identifiers some tokenizers return are dropped before generation, which a causal model's forward pass does not accept and which had left three of the second cohort's candidates without output (Appendix E). For the thirteen models evaluated before the amendment the line is a no-op: none of their tokenizers returns that field, each of the three affected ones does, and the amended harness reproduces the recorded 60-request screen of one second-cohort model row for row.

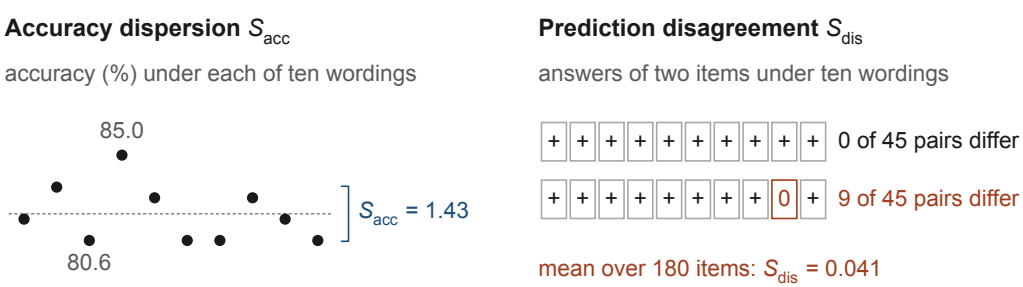


Figure 3: The two sensitivity features on one development cell, Mistral 7B on the Amazon sentiment source split. Left: source accuracy under each of the ten wordings, with their mean (dashed) and one standard deviation either side, which is S_{acc} for this cell. Right: the answers two items receive under the ten wordings (+ positive, 0 neutral). Each item contributes the share of its 45 wording pairs that return different labels: none for the first item, 9 of 45 for the second, where one wording dissents. Their mean over the 180 evaluation items is S_{dis}; it needs no labels.

Table 4: The frozen measurement setup (`PIPELINE_FROZEN.txt`, items 3 to 5), with the file that implements each item.

Item	Value
Library, precision	Hugging Face `transformers` 4.52.4; weights loaded in fp16 (`hf_backend.py`)
Prompt	each model's own chat template, one user turn
Decoding	greedy, thinking disabled in the chat template, at most 128 new tokens
Requests	one item per request, batched with left padding (at most 48 requests or 12,000 tokens); no retries
Context	a request whose prompt and 128 new tokens exceed the model's context is not generated and scores invalid; a shift where this exceeds 5% of requests is unavailable for that model
Wordings	ten per task, one canonical (`prompt_variants.json`)
Data	BOSS splits (Yuan et al., 2023) at revision `974d49e`, official test splits; SNLI test (`stanfordnlp/snli`) at `cdb5c3d`, its 176 rows without a gold label removed first
Item selection	200 items per split, drawn with seed 32001 for every split; each split's item list and seed are in `*.manifest.json`, SNLI drawn by `make_snli_split.py`
Items scored	items 21 to 200 (180) of every split; the first 20 are screening items for every model
Rule R1 (registered)	strip at most one enclosing Markdown fence, parse JSON, take the entry for the item's identifier; its label, stripped and lowercased, must equal an allowed label; otherwise, or if the response reached 128 tokens, invalid, kept in the denominator and its own label value in S_{dis} (`extract.py`)

What three families can resolve. A sign-flip test over k units cannot go below a one-sided p of $1/2^k$: 0.0156 at $k = 6$, 0.03125 at $k = 5$, 0.125 at $k = 3$. With the level unadjusted, six is the smallest cohort that can reject at 0.025, and only if the feature helps in every family: a floor, not a power calculation.

C THE EVALUATED GRID

Two development candidates failed the screen of Section 4.1, one returning no valid screening response and one 28%.

D THE HELD-BACK SHIFT, DESCRIPTIVE

Registered as descriptive, MNLI to SNLI, target side only: the registered fit on the 49 development cells, none of them on SNLI, predicts it with its SNLI indicator at zero. The baseline does not predict on this shift: its mean absolute error, 19.2 points for the development models and 19.3 for the held-out ones, exceeds a training-mean predictor's 15.2 and 16.2, and its R^2 is negative in both blocks. By the rule of Section 3.2 that a relative improvement is read only where the baseline's own

Table 5: The fifteen evaluated models, each quantity averaged over the model's cells; accuracy, dispersion and degradation in points; Invalid: percent of invalid responses, source/target side. Four models exceed the context limit on one inference shift in more than 5% of requests and so contribute six cells. Snapshot: the Hugging Face commit of the loaded weights, recorded in every output row and, for held-out models, pinned in advance (`holdout_manifest.json`, `MANIFEST.json`, `cohort3_manifest.json`).

Repository	Family	A_{src}	S_{acc}	S_{dis}	D	Invalid	Snapshot
Development pool, 49 cells							
`Qwen/Qwen3-1.7B`	qwen3	73.4	3.10	0.124	17.9	0.1/0.1	`70d244c`
`Qwen/Qwen3-4B`	qwen3	76.6	1.92	0.056	10.7	0.3/0.4	`1cfa9a7`
`ibm-granite/granite-3.3-2b-instruct`	granite3	78.5	5.36	0.128	21.7	3.7/5.7	`707f574`
`tiiuae/Falcon3-3B-Instruct`	falcon3	74.7	9.20	0.167	20.5	5.2/10.0	`411bb94`
`tiiuae/Falcon3-7B-Instruct`	falcon3	85.8	1.64	0.035	19.7	0.0/0.2	`1e57a0e`
`microsoft/Phi-4-mini-instruct`	phi4	81.2	1.42	0.046	17.0	0.1/0.1	`cfbefac`
`mistralai/Mistral-7B-Instruct-v0.3`	mistral	75.4	2.14	0.077	13.0	0.4/0.4	`c170c70`
First held-out cohort, 19 cells							
`allenai/OLMo-2-1124-7B-Instruct`	olmo2	80.7	5.46	0.101	18.0	1.2/1.8	`470b1fb`
`HuggingFaceTB/SmolLM2-1.7B-Instruct`	smollm2	57.0	3.94	0.161	2.6	0.1/3.9	`31b70e2`
`01-ai/Yi-1.5-9B-Chat`	yi1.5	76.3	3.33	0.109	9.6	4.2/5.7	`1a0fc69`
Second held-out cohort, registered after the first result, 20 cells							
`kakaocorp/kanana-1.5-8b-instruct-2505`	kanana	77.5	2.00	0.066	13.3	0.0/0.0	`c963a5f`
`yandex/YandexGPT-5-Lite-8B-instruct`	yandexgpt	79.1	1.26	0.034	12.0	0.2/0.2	`b556811`
`Zyphra/Zamba2-7B-Instruct`	zamba2	79.9	1.63	0.051	15.5	1.8/0.9	`3146688`
Third held-out cohort, pre-specified after the second result, 13 cells							
`K-intelligence/Midm-2.0-Mini-Instruct`	midm	69.0	4.04	0.113	10.8	0.0/0.1	`383eb22`
`llm-jp/llm-jp-3-7.2b-instruct3`	llm-jp	56.8	7.08	0.219	5.5	13.0/8.4	`cdd4c7f`

held-out R^2 is positive, Table 6 is therefore only recorded: disagreement lowers error for four of five development families and dispersion for one, and neither lowers it for any held-out family. Its degradations are small or negative for some families, so it is a weak shift as a test. In development, withholding one shift at a time its R^2 falls to 0.14 (Table 1), and withholding a task group leaves it no learned intercept (R^2 of -1.02), so that scheme is not reported further.

Table 6: The held-back shift, MNLI to SNLI, per family (descriptive, no p-value, blocks not pooled). D: observed degradation, mean over the family's models, in points. Baseline error: absolute error of the registered fit on the 49 development cells, which contain no SNLI cell. Change: reduction in that error when the feature is added, positive when error fell.

Family	Cells	D	Baseline error	Change, $+S_{\text{acc}}$	Change, $+S_{\text{dis}}$
Development families: the shift is new					
Falcon3	2	-6.2	31.65	-0.12	+0.25
Granite 3.3	1	1.8	17.84	-0.04	-2.18
Mistral	1	17.1	1.08	+0.85	+0.82
Phi-4	1	-4.5	32.26	-0.08	+0.64
Qwen3	2	5.8	9.84	-0.14	+1.25
Families where error fell				1/5	4/5
Held-out families: family and shift both new, not attributed to either					
OLMo 2	1	3.1	21.16	-2.03	-2.16
SmolLM2	1	3.1	15.88	-5.27	-5.54
Yi 1.5	1	-3.3	21.00	-1.25	-5.00
Families where error fell				0/3	0/3

E THE SECOND HELD-OUT COHORT

Registered on 24 September (Section 3.3): candidate order, screen, frozen predictor, an outcome rule and the no-pooling clause. The rule reads, per feature, the pooled change in MAE when the feature is added: a rise of at least 0.30 points returns "A second, result-blind cohort replicates the harm.", a change strictly between ± 0.30 returns "no practical effect in the new cohort", and a fall of at

least 0.30 returns "held-out evidence is mixed", never "helps", with the families helped stated; with two evaluated families it allows per-family effects only, and with one or none the screen outcome only. Candidates were ordered before any download, one model per family, and screened one at a time until three passed (Table 7). The three without output failed because the frozen harness passes `token_type_ids` to models that do not accept it, and four of the five candidates not reached are gated repositories we could not access. After the harness amendment of 25 September (Appendix B) the eight remaining candidates were screened for the third cohort in the same order, and its pre-specified rule, written after both earlier cohorts' directions were known, reads a pooled change of at least 0.30 points in either direction at three families and at two returns the per-family differences and the pooled change as a description; the cohort has two families (Section 5.5). A registered retry of those three, had fewer than three passed, was not triggered (amendment A1). On the 20 cells, fit on the 49 development cells only, the falls in pooled error are 0.68 points for dispersion and 1.82 for disagreement from unrounded MAE, each in 3 of 3 families (exact sign-flip p 0.125 one-sided, 0.25 two-sided, resolution-limited), so the rule returns mixed for both; an independent recomputation from the raw responses reproduces this. Amendment A2, registered before any cohort output was opened, changed when one run was computed, not what: the slow Zamba2 run was split into parallel parts of the frozen harness, its first 2,400 rows kept.

Table 7: The second and third cohorts' candidates in registered order (repositories and revisions in `MANIFEST.json`), screened one at a time until 3 passed: 9 screened, 3 passed, 3 failed the format criterion, 3 produced no output. In parentheses: share of the 60 screening requests with a valid response. Gated: a gated repository we could not access. A2: run split under amendment A2 (Appendix E). After the harness amendment of 25 September (Appendix B) the 8 remaining candidates were screened for the third cohort in the same order: 2 passed, 2 failed the format criterion, 4 could not be fetched.

Order	Model	Status
1	`Tri-7B`	no output under the frozen harness; screened again after the amendment: failed format (0.000)
2	`A.X-3.1-Light`	failed format (0.633)
3	`Midm-2.0-Mini-Instruct`	no output under the frozen harness; screened again after the amendment: passed (0.983), 7 cells
4	`kanana-1.5-8b-instruct-2505`	passed (1.000), 7 cells
5	`YandexGPT-5-Lite-8B-instruct`	passed (0.983), 7 cells
6	`sarashina2.2-3b-instruct-v0.1`	failed format (0.850)
7	`llm-jp-3-7.2b-instruct3`	no output under the frozen harness; screened again after the amendment: passed (0.917), 6 cells
8	`Minerva-7B-instruct-v1.0`	failed format (0.317)
9	`Zamba2-7B-Instruct`	passed (1.000), 6 cells, A2
10	`salamandra-7b-instruct`	not reached; after the amendment: failed format (0.000)
11	`HyperCLOVAX-SEED-Text-Instruct-1.5B`	not reached (gated); after the amendment: could not be fetched (gated)
12	`Velvet-2B`	not reached (gated); after the amendment: could not be fetched (gated)
13	`c4ai-command-r7b-12-2024`	not reached (gated); after the amendment: could not be fetched (gated)
14	`EuroLLM-9B-Instruct`	not reached (gated); after the amendment: could not be fetched (gated)

F THE LABEL-FREE FEATURE IN TWO PARTS (PRE-SPECIFIED)

Table 8 splits disagreement exactly into two parts, each entered alone beside source accuracy. The pre-specified rule for this split reads formatting-carried if the invalid part averages at least half of the statistic and C_{valid} moves error by at most 0.10 points, semantic if C_{valid} alone recovers at least 80% of the gain of S_{dis}, and mixed otherwise; it returns mixed, the invalid part averaging 15.0%. Descriptive: C_{invalid} alone recovers 78.6% of the 0.73-point development gain of S_{dis}, C_{valid} none

of it, and neither lowers the first cohort's error. On the second cohort they recover 34.3% and 39.7% of the pooled gain; there C_{invalid} is exactly 0 in 13 of 20 cells, which carry 86% of its pooled drop, so that drop comes mostly from the refitted development coefficients, not from the cells' own values.

Table 8: The label-free feature split exactly into its two parts (Appendix F). C_{valid}: share of wording pairs giving two different valid labels; C_{invalid}: share where exactly one response is invalid; $S_{\text{dis}} = C_{\text{valid}} + C_{\text{invalid}}$ on every cell. Each enters alone beside source accuracy. MAE in points, with the change against the baseline and units helped; development: one family out; cohorts: one fit on the 49 development cells, the second cohort's columns descriptive. Mean share of C_{invalid} in S_{dis} over the 49 development cells: 15.0%. Pre-specified outcome: mixed.

	Development		First cohort		Second cohort	
Source-side predictors	MAE	Change	MAE	Change	MAE	Change
A_{src} only (baseline)	6.15		6.58		5.21	
$+S_{\text{dis}}$ (registered feature)	5.42	+11.8%, 4/5	6.97	−6.0%, 1/3	3.38	+35.0%, 3/3
$+C_{\text{valid}}$	6.15	−0.1%, 3/5	8.18	−24.4%, 1/3	4.48	+13.9%, 3/3
$+C_{\text{invalid}}$	5.58	+9.3%, 5/5	6.81	−3.5%, 1/3	4.58	+12.0%, 3/3

G PAIR DELETIONS (PRE-SPECIFIED)

Each of the 21 pairs of development models is removed in turn, the registered external fit is repeated on the other five, and the same 19 first-cohort cells are predicted (Figure 4); the both-Qwen3 pair is the Qwen3 deletion of Figure 2. For disagreement the pre-specified rule returns Qwen3-specific: 4 of 21 pairs reverse, all containing Qwen3 1.7B, none of the 10 without a Qwen3 model. For dispersion it returns mixed: 3 of 21 reverse, 1 of the 10 without a Qwen3 model (Falcon3 3B with Granite 3.3 2B). The baseline's held-out R^2 stays between 0.57 and 0.79, inside the pre-specified band (0.55 to 0.80) beyond which a fit is flagged; equal family weights read "robust to equal family weights".

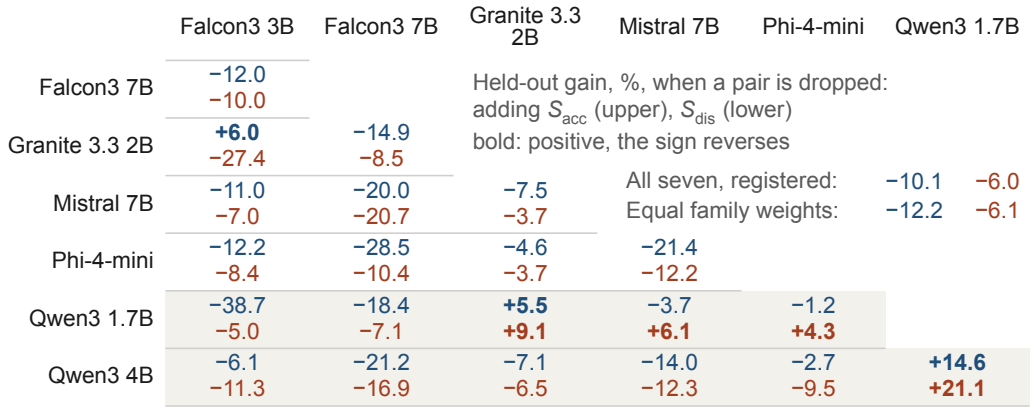


Figure 4: First-cohort held-out gain, in percent, from adding S_{acc} (blue, upper) or S_{dis} (rust, lower), positive when error fell, with the registered fit repeated without each pair of development models, one cell per pair; the fits on all seven are printed top right. Bold: a positive gain, where the sign reverses. Shaded rows: every pair containing a Qwen3 model. Pre-specified; the pairs overlap and the spread is not a confidence interval.

H THE SCORING RULE AND ITS EXTRACTION AUDIT (PRE-SPECIFIED)

The alternative rule of Section 6.2 recovers a label only when a rejected response identifies exactly one allowed label for the requested item unambiguously: the first fenced block if any, else the first brace-complete JSON object, at most one bracket-closing repair, an exact item-identifier match; it never changes an accepted label and drops no response. Over the development pool and the first cohort it recovers 1,790 of the 5,225 rejected responses, 34.3%, which is coverage, not correctness:

Table 9: Scoring rule (pre-specified). The alternative rule rescores only responses the registered rule rejected. Source side only: features rescored, outcome kept at the registered degradation; end to end: both rescored. MAE in points, with the change against that row's own baseline, positive when error fell, and units helped.

	Baseline	$+S_{\mathrm{acc}}$	$+S_{\mathrm{dis}}$
Development, one family out			
Registered rule	6.15	5.64 (+8.3%, 5/5)	5.42 (+11.8%, 4/5)
Alternative, source side only	5.76	5.51 (+4.4%, 4/5)	6.19 (−7.4%, 3/5)
Alternative, end to end	5.69	5.48 (+3.8%, 3/5)	6.43 (−12.9%, 3/5)
First held-out cohort			
Registered rule	6.58	7.24 (−10.1%, 1/3)	6.97 (−6.0%, 1/3)
Alternative, source side only	6.80	7.42 (−9.0%, 1/3)	6.73 (+1.1%, 1/3)
Alternative, end to end	7.36	7.92 (−7.6%, 2/3)	7.23 (+1.7%, 1/3)
Second held-out cohort			
Registered rule	5.21	4.52 (+13.1%, 3/3)	3.38 (+35.0%, 3/3)
Alternative, source side only	4.92	4.48 (+8.9%, 3/3)	3.42 (+30.6%, 3/3)
Alternative, end to end	5.11	4.71 (+7.9%, 3/3)	3.56 (+30.4%, 3/3)

recovered labels match the gold label 57.5% of the time, against 66.5% for responses the registered rule accepted and 37.7% chance on the same items. For extraction fidelity, a pre-specified audit, fixed before any response was read, gave ten recovered responses per parsing pattern, sampled at random with a floor of one per model, to a reader, Claude (Sonnet 5), shown only the model name, item identifier and raw response. It agreed with the rule's label on 10 of 10 in each of the three patterns, 30 of 30 (exact binomial 95% interval 88.4 to 100%, patterns weighted equally). This is agreement between two automated readers, not a human audit; the 3,435 responses invalid under both rules are outside it. In the second cohort the rule rescues 0 of 2 rejected responses for Kanana, 0 of 38 for YandexGPT and 92 of 1,888 for Zamba2, whose rejections fall mostly (1,721) on the shift excluded from its cells; the other 167 are the invalid rates of Table 5. In the first cohort it rescues 813 of 1,281 for Yi 1.5.

I THE PROBIT CALIBRATION (PRE-SPECIFIED)

Specified before it was run, as the only stronger calibration tried. With a the source accuracy and $t = a - D$ the target accuracy, both in points and clipped to $[1, 99]$, the fit regresses $\Phi^{-1}(t/100)$ on $\Phi^{-1}(a/100)$ and the task and shift indicators with the frozen ridge estimator and penalty search, adding S_{acc} or S_{dis} for the augmented fits; predicted degradation, $a - 100\,\Phi(\hat{\eta})$, is scored in points as in Table 1. With d_1 and d_2 the change in MAE from adding S_{dis} on the two cohorts (positive: error rose), the rule fixed in advance calls the opposite directions persistent if $d_1 > 0$, $d_2 < 0$ and both are at least 0.25 points in size. Here $d_1 = +0.28$ and $d_2 = -1.69$.

J CONTROLS (PRE-SPECIFIED)

Placebo (C1). The permutation unit is the (model, source dataset) value, shared by every shift leaving that source, and the stratum is (block, source dataset). A replicate draws a uniformly random permutation of the models in each stratum of the development and held-out blocks and gives each model another's feature value on that source in all its cells there, with the same indices for S_{acc} and S_{dis}; nothing else is touched. The reference keeps the feature's per-source marginals and breaks its association with the model, hence with source accuracy, invalid rate and lineage: a permutation reference under this fitted procedure, not a test of information beyond accuracy. The gains are recomputed through the frozen pipeline fit on the permuted development block, $N = 4000$, seed 20260925; with G the placebo gains and g the observed one, $p_{\mathrm{help}} = (1 + \#\{G \geq g\})/(N + 1)$ and $p_{\mathrm{harm}} = (1 + \#\{G \leq g\})/(N + 1)$, read at 0.05 in the direction of g. Before the draws: the identity permutation returns every registered gain to 10^{-12}; a planted informative column (each model's mean baseline residual per source), left unpermuted while the draws permute it, is above the reference at $N = 200$ with $p = 0.005$, that is, no draw reached it; a permutation across source strata is refused. For the third cohort the same procedure runs on the development and third-cohort blocks

with seed 20260926 and $N = 4000$; the development and held-out blocks are permuted together, so no reference here is a held-out-only permutation test, and with two models per source the third cohort has only 8 distinct held-out patterns: disagreement's fall exceeds it (placebo mean -0.13, $p = 0.0020$) and dispersion's lies within it (placebo mean -0.14, $p = 0.062$).

Compliance (C2). Arms under the same folds and estimator: the source-side invalid-response rate alone; each feature residualised on it by least squares with intercept fitted on training cells only (the four training families inside each development fold; the 49 development cells for a held-out cohort); and a log-parameter-count arm, dropped because the counts exist for the held-out cohorts and not for the seven development models. The protocol's labels, first match: "no gain to attribute" below a 0.15-point gain; "more than compliance" if the residual keeps at least half and at least 0.15 points while the invalid rate alone falls short of the gain by more than 0.15; "carried by compliance" if the residual keeps less than half; "mixed" otherwise. It is a residualisation sensitivity analysis, not an attribution. A block with a nonzero invalid rate in fewer than half its cells is degenerate: its numbers are printed, its verdict not read. The held-out residualised-disagreement gains, -1.00, -0.47 and -9.56, carry no verdict: the first cohort has no gain to attribute and the second and third are degenerate (7 of 20 and 6 of 13 cells nonzero; in the third the invalid rate alone gives $+2.28$ points and the residualised dispersion $+0.96$), and none is evidence about what disagreement carries. A descriptive diagnostic of the third cohort's -9.56, not a new arm: the residualised column is more collinear with source accuracy in the development design (correlation of the standardised columns -0.84, against -0.61 for raw disagreement); all six llm-jp-3 residuals lie at or below the development residual minimum (-0.104 to -0.046 against -0.043 to $+0.112$), with its cells far below the development source-accuracy range, and there the two-feature ridge fit is unstable (llm-jp-3 error 39.09 points against 12.47 for the baseline; without its worst cell the pooled residualised error is still 16.14 against 10.29), while for Midm, whose residuals lie inside the range, it is lower (2.94 against 8.01). The pooled figure (19.63 against 10.07) is a fit extrapolating outside its training support under collinearity, not a finding about compliance or about sensitivity. The invalid rate correlates with S_{acc} at $r = 0.86$ and with S_{dis} at 0.82 over the 21 development model-source units (0.74 and 0.55 over the thirteen models of the development pool and the first two cohorts).

Wording resampling (C3). All $\binom{9}{6} = 84$ seven-wording subsets containing the canonical one; for each, A_{src}, S_{acc}, S_{dis} and D are recomputed from the raw responses of the thirteen models of the development pool and the first two cohorts, and the full-set path reproduces the registered cells and gains. A direction is stable if at least 76 of 84 subsets keep the full-set sign; bands are 5th and 95th percentiles over subsets that overlap and are not independent. The direction holds in 66 of 84 subsets for dispersion in development (band -0.83 to $+0.62$ points) and in 76 for disagreement (-0.18 to $+1.13$); in the first cohort in 84 for dispersion (-1.93 to -0.22) and 63 for disagreement (-1.35 to $+0.49$); in the second in 83 for dispersion ($+0.37$ to $+1.00$) and 84 for disagreement ($+1.46$ to $+1.89$).

Six- and eight-family summaries (C4; descriptive, not a test). For the per-family changes d_f of Figure 1, per feature: the mean, the t-interval $\bar{d} \pm t_{0.975,n-1}\, s/\sqrt{n}$ with s the sample standard deviation, and the 2.5th and 97.5th percentiles of the mean over 10000 family bootstrap resamples (seed 20260925); no test was registered or run over the families. Over the six families of the first two cohorts the change averages $+0.05$ points for dispersion and $+0.71$ for disagreement (t-intervals -1.43 to $+1.53$ and -0.73 to $+2.14$, bootstrap -1.10 to $+0.92$ and -0.26 to $+1.72$, 4 of 6 positive for each); over the eight families with the third cohort added, $+0.24$ and $+1.30$ (t-intervals -0.85 to $+1.32$ and -0.06 to $+2.66$, bootstrap -0.71 to $+0.97$ and $+0.24$ to $+2.37$, 6 of 8 positive for each), the families having been added in three sequential cohorts.

Support (C5; descriptive). Table 10. Descriptively, with SmolLM2's seven cells set aside and the fit unchanged, the first cohort's pooled error is 5.31 points for the baseline, 4.87 with dispersion ($+8.4\%$) and 5.33 with disagreement (-0.3%). Over the six families of the first two cohorts the Spearman rank correlation of the per-family change with distance outside the source-accuracy range is -0.65 for both features, and with distance outside the degradation range -0.17 for dispersion and -0.68 for disagreement; over the eight families with the third cohort added, -0.03 and $+0.30$ with the source-accuracy distance and $+0.11$ and -0.22 with the degradation distance (descriptions of six and eight points, no p-value).

Table 10: Support (descriptive, $n = 8$; the third cohort's two families were added after the six): model-level mean source accuracy and degradation; distance outside the development model-level range of each (73.4 to 85.8 and 10.7 to 21.7; zero inside); cells outside the development cell-level bounding box; nearest development model in standardised distance; change in held-out MAE per feature.

		Mean		Outside range		Cells	Nearest	Change	
Family	Cohort	A_{src}	D	A_{src}	D	outside	development model	$+S_{\mathrm{acc}}$	$+S_{\mathrm{dis}}$
OLMo 2 7B	1	80.7	18.0	0.0	0.0	0/6	Phi-4-mini	-0.37	-0.31
SmolLM2 1.7B	1	57.0	2.6	16.4	8.1	4/7	Qwen3 4B	-2.56	-1.04
Yi 1.5 9B	1	76.3	9.6	0.0	1.1	0/6	Qwen3 4B	$+1.26$	$+0.27$
Kanana 1.5 8B	2	77.5	13.3	0.0	0.0	1/7	Mistral 7B	$+0.74$	$+2.19$
YandexGPT 5 Lite 8B	2	79.1	12.0	0.0	0.0	0/7	Qwen3 4B	$+1.06$	$+2.40$
Zamba2 7B	2	79.9	15.5	0.0	0.0	0/6	Phi-4-mini	$+0.18$	$+0.73$
llm-jp-3 7.2B	3	56.8	5.5	16.6	5.2	3/6	Mistral 7B	$+1.52$	$+3.71$
Midm 2.0 Mini	3	69.0	10.8	4.4	0.0	1/7	Mistral 7B	$+0.07$	$+2.43$

Post hoc checks (C6; descriptive, no rule in advance). Table 11. Top: on the third cohort the two parts of the label-free feature enter alone beside source accuracy. C_{invalid} lowers error by 15.8% and C_{valid} raises it by 50.3%, while S_{dis} lowers it by 30.0%, so the parts do not add. Bottom: with the development fit unchanged, the feature values are reassigned among the held-out models within each source, over every arrangement (216 for three models per source, 8 for two). Dispersion's gain on the second cohort, $+0.68$ points, is at the mean of that reference ($+0.70$); disagreement's, $+1.82$, lies above its mean ($+1.36$) and inside its range (exact $p = 0.13$); on the third cohort disagreement's gain is the largest of the 8 arrangements ($p = 0.125$, resolution-limited: the smallest attainable), against a mean of $+0.80$. This reference keeps the fit and breaks only which held-out model carries which value, whereas the placebo of C1 refits on the permuted development block; they are different references. A planted column carrying each cell's own model-and-source mean degradation is rejected at $p = 0.0093$ in the first and second cohorts.

Table 11: Post hoc, descriptive (C6). Top: on the third cohort the two parts of the label-free feature (Appendix F) enter alone beside source accuracy; MAE in points, change against the baseline, units helped; one fit on the 49 development cells. Bottom: exact held-out-only permutation. The development fit is unchanged and the feature values are reassigned among the held-out models within each source, over all N arrangements; gain in points of MAE, mean and 5th to 95th percentile of the arrangements, exact one-sided p.

Third cohort	MAE	Change	Helped
A_{src} only (baseline)	10.07		
$+S_{\mathrm{dis}}$ (registered feature)	7.05	$+30.0\%$	2/2
$+C_{\mathrm{valid}}$	15.13	-50.3%	1/2
$+C_{\mathrm{invalid}}$	8.48	$+15.8\%$	2/2

Cohort	Feature	N	Gain	Mean	5th to 95th	p
First	S_{acc}	216	-0.66	-0.08	$[-0.77, +0.95]$	0.8889
First	S_{dis}	216	-0.39	-1.38	$[-3.66, +0.39]$	0.2407
Second	S_{acc}	216	$+0.68$	$+0.70$	$[+0.59, +0.82]$	0.6667
Second	S_{dis}	216	$+1.82$	$+1.36$	$[+0.80, +1.97]$	0.1296
Third	S_{acc}	8	$+0.74$	$+0.54$	$[+0.05, +1.03]$	0.5
Third	S_{dis}	8	$+3.02$	$+0.80$	$[-1.27, +2.86]$	0.125