WHAT SHOULD BLACK-BOX LLM CONFIDENCE PREDICT?
ESTIMATING THE CORRECTNESS OF A FIXED ANSWER FROM MULTIPLE SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Black-box confidence methods score an answer by asking the model, resampling it, perturbing the input, or consulting other models, but need not score the same object. We freeze each question–answer pair before measurement and ask what survives new models and questions. Signals derived from the target itself, self-report and resampling, mostly measure commitment, including commitment to stable errors. Other models supply external corroboration. Under simultaneous model and question holdout, a small head that fuses self-report with peer evidence beats retrained text encoders and a released correctness model, and a frozen text encoder given the peer measurements nearly recovers the fusion. Asking the same peers to judge the frozen answer reads more of that evidence than counting their agreement on five of six datasets. A monotone environment link transfers the ranking while the probability scale stays local; peer value is governed by the panel's separation between right and wrong target answers; and the correctness of a unanimous consensus leaves the unlabeled agreement law unchanged. Separation predicts that stale agreement ages when an upgrade erodes it. Three frontier models showed it, the second time in a preregistered replication: the same previous-generation peers weakened whether counted or asked to judge, refreshed peers restored both readouts, above self-report at a moderate gap and indistinguishable from it at a wide one, and pooled stale peers diluted it. Freeze the answer, read capable peers as judges, refresh them when the target moves, audit the scale locally, and never equate unanimity with truth.

1 WHAT SHOULD BLACK-BOX CONFIDENCE PREDICT?

Deployments change faster than the calibrators built for them. Answer models are replaced across generations, and labels from the newly deployed model are scarcest at exactly the moment its confidence matters most. A confidence method that must be retrained for every model and dataset pair is costly to keep current, so a deployment needs to know which parts of a confidence score survive such changes, which parts break, and what a repair costs.

The object being scored must be fixed first. Suppose a black-box model M receives a question q_i and returns an answer a_i, and a downstream system must decide whether to use the answer, verify it, or abstain. The quantity needed for this decision is

$$\eta_i = \mathbb{P}_{\mathcal{D}}\left(Y_i = 1 \mid \text{available observations about } q_i, a_i\right), \tag{1}$$

where $Y_i = 1$ means that the answer is correct and $\mathcal{D}$ is the deployment distribution. Throughout, *confidence* is a score attached to one fixed answer, *uncertainty* is spread in the answer distribution before an answer is fixed, and *calibration* is agreement between probabilities and observed frequencies. A score can rank answers well without being calibrated, and calibration cannot rescue a score with no correctness information.

We ask three questions: what black-box signals measure on one fixed answer; which properties survive unseen models and questions; and the target-label cost of maintaining them. Our contributions are:

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.

1. **One fixed answer, scored under simultaneous model and question holdout.** We freeze each question–answer pair and predict its correctness with supervision that never sees the deployed model or the evaluated questions. Freezing the answer makes confidence instruments comparable, and the fixed answer separates commitment from corroboration. Target-derived self-report and resampling primarily measure commitment, while peers contribute external corroboration.
2. **Peer evidence carries the transferable signal, and how it is read matters.** Under that holdout, a small head that fuses self-report with peer agreement outperforms exactly retrained text encoders on four panels, and on the three 512-question panels also a released correctness model and a five-family perturbation composite. A frozen encoder given the peer measurements nearly recovers the fusion, so the tested text predictors lack the information and not the capacity. The same peers can be read two ways: counting their agreement needs no verifier and, on letter-answer datasets, no call beyond their answers; asking them to judge the frozen answer reads more on five of six datasets, including a panel that is right only half the time; one small head takes both.
3. **Peer evidence ages, whatever the readout; the target's own report does not.** A monotone environment link transfers the ranking while the probability scale stays local, and peer value is governed by the panel's separation between right and wrong target answers. This account predicts that stale agreement hurts when a model upgrade erodes that separation. A first generation change showed it, a preregistered replication confirmed it, and the same peers weakened as judges. Refreshed peers restore both readouts, above self-report at a moderate gap and indistinguishable from it at a wide one, while self-report keeps working. The same account prices the labels each deployment decision needs.

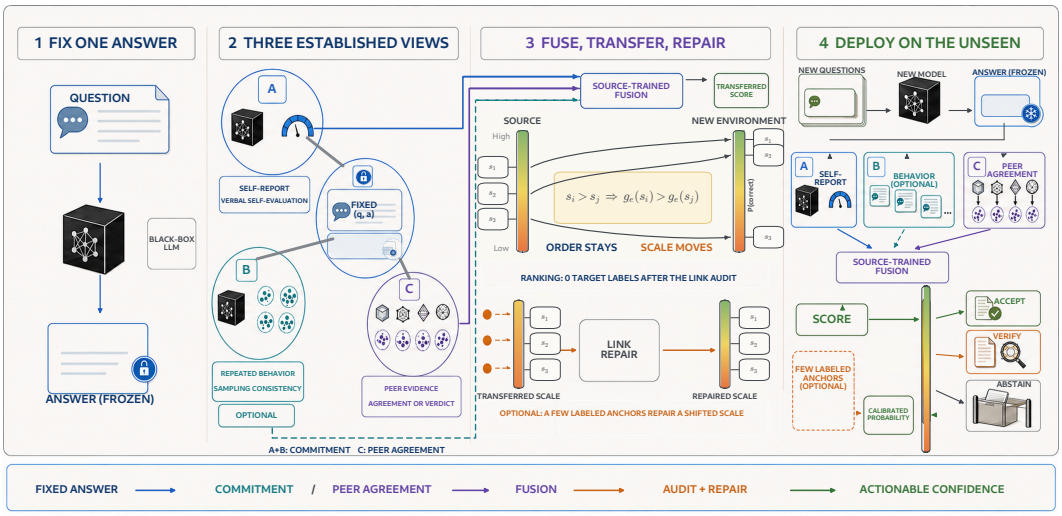

Figure 1: The fixed-answer framework. Freeze each question–answer pair before measurement; instruments may read but never replace it. The views place established signal families on one common object: A and B mainly measure commitment, C peer evidence (their agreement or their verdict), and B is optional. Under a monotone link the ranking transfers; the probability scale stays environment-local, audited with labels and, where the affine form holds, repaired with about twenty anchors. The score routes each answer to accept (use it), verify it, or abstain.

Figure 1 summarizes the argument; Sections 3–5 establish transfer, its mechanism, and the predicted failure, and Section 6 prices the labels each decision needs. Appendix A opens with a table mapping each claim to its decisive test, and Appendix T lists the open questions.

2 A FIXED ANSWER REVEALS COMMITMENT AND CORROBORATION

Black-box confidence methods ask the target directly (Kadavath et al., 2022; Tian et al., 2023; Xiong et al., 2024), resample it (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024), perturb

the input (Pedapati et al., 2025), or consult other models (Shrivastava et al., 2023; Hamidieh et al., 2026) (Appendix B reviews what each family observes). These procedures are often compared as if they measured one quantity, but without controlling the answer they can change both the instrument and the object being scored. We instead generate one neutral answer per question, freeze the pair, and require every measurement and correctness label to refer to that object.

2.1 One object, three views

View A asks the target, through a fixed set of prompts, for the probability that its frozen answer is correct. View B reruns the exact answer prompt at several temperatures and measures support and semantic entropy around that answer. View C asks provider-diverse peers the same question. A gold-blind judge clusters their answers, from which we record the fraction of peers agreeing with the target and three partition-shape statistics (largest-cluster share, cluster entropy, unique-answer fraction). Judge swaps are audited in Appendix D. The same peers can instead be shown the frozen answer and asked how likely it is correct. We call that view J and compare it with C in Section 3. The peer view below always means C.

2.2 Commitment is not correctness

A one-target SimpleQA pilot with 128 questions (Appendix C) reveals why these views should not be averaged. Self-report and repeated sampling share one detected direction: a confident, stable, concentrated answer. Both signals are generated by the target itself, so a consistently reproduced answer scores high on both, and that consistency need not track truth, since both remain high on some wrong answers. We call this shared signal *commitment*. Correctness information absent from the target's own commitment must come from elsewhere, and peer corroboration supplies it. Full-scale resampling and paraphrase controls reach the same conclusion (Appendix G).

2.3 A minimal estimator and a transfer test

Our primary estimator concatenates A and C, standardizes the features, and fits regularized logistic regression on source labels, so the predictor has at most eleven parameters and involves no target-model fine-tuning. View B is optional, since given the peer view it adds little except on GSM8K (instruments, a one-prompt variant, and costs: Appendices E and F).

Every primary evaluation holds out both the target answer model and a question block. This double holdout removes the shortcuts that random splits preserve. The scaler and predictor are refit inside each source split, and no target-model or held-out-question label enters training. Brier score and log loss evaluate the probability, while AUROC evaluates ranking.

3 Peer evidence transfers; a verdict reads more of it than a count

The transfer test holds out a new answer model and question block on four panels. SimpleQA contains 128 questions and eight answer models; GSM8K, TriviaQA, and MMLU-Pro contain roughly 512 questions and five provider-diverse models each (Wei et al., 2024; Cobbe et al., 2021; Joshi et al., 2017; Wang et al., 2024). We retrain Calib-text (Devlin et al., 2019), two APRICOT variants (Ulmer et al., 2024; He et al., 2023), and Calib-n-joint (Xia et al., 2025), the last adapted to score an unseen model (Appendix F), inside the same source split. Two zero-target-label comparators are the released Generalized Correctness Model (Xiao et al., 2025), an 8B model trained on eight source models' correctness histories and applied zero-shot, and the five-family black-box perturbation composite of Pedapati et al. (2025). All methods score the same fixed answers.

Method	SimpleQA AUC↑	SimpleQA Br↓	GSM8K AUC↑	GSM8K Br↓	TriviaQA AUC↑	TriviaQA Br↓	MMLU-Pro AUC↑	MMLU-Pro Br↓
A+C (minimal)	**.823**	**.136**	.938	.113	.933	.038	.859	.128
A+B+C (+8 target resamples)	N/A	N/A	**.968**	**.070**	**.938**	**.037**	**.862**	**.125**
C only	.797	.138	.837	.147	.922	.040	.836	.143
A only	.661	.166	.804	.127	.774	.067	.756	.175
B only (8 target resamples)	N/A	N/A	.805	.163	.726	.083	.650	.189
Perturbation composite (five families)	N/A	N/A	.901	.123	.848	.076	.750	.186
Correctness model (8B, zero-shot)	N/A	N/A	.808	.180	.863	.088	.714	.224
Best encoder by AUROC	.581	.261	.630	.258	.671	.100	.669	.212
Source prevalence	.333	.179	.299	.275	.378	.099	.345	.234

Table 1: **Peer corroboration carries transfer; target resampling adds most on GSM8K.** All methods face simultaneous held-out answer-model and question-block shift. AUC is AUROC; Br is Brier score. A+C is the minimal fusion; A+B+C adds eight target resamples; N/A marks a configuration not evaluated on SimpleQA under this protocol. "Best encoder by AUROC" reports that encoder's complete row. The perturbation composite fuses the five implementable families of Pedapati et al. (2025), scored on the rows with valid perturbations; the correctness model is the released Xiao et al. (2025) adapter, applied without our labels. Protocol details, full log loss, every encoder, and uncertainty are in Table 12 and Appendices G, E, and H.

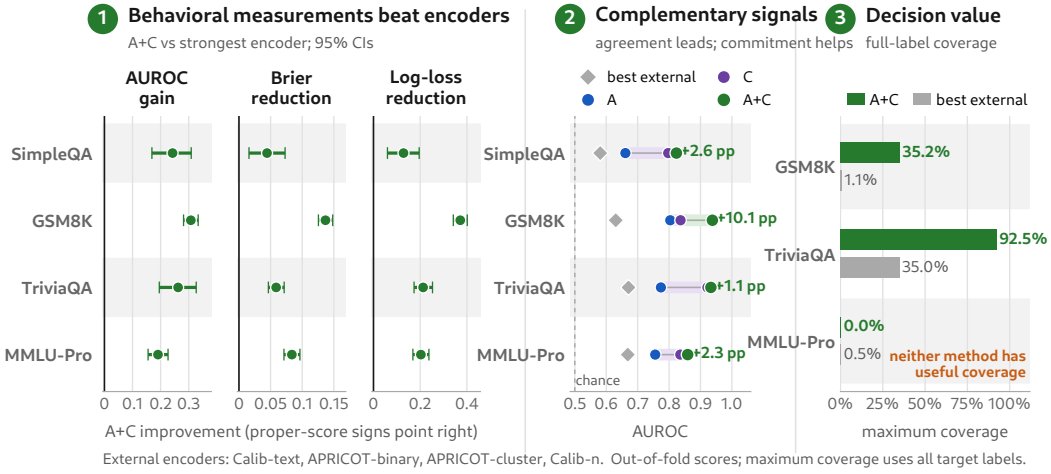

Figure 2: **Transfer follows complementary information.** Panel 1: improvement of the A+C fusion (at most 11 parameters) over the strongest retrained encoder, selected per dataset and metric (the two additional comparators are in Table 1); proper-score signs point right and intervals are 95% question-cluster bootstrap CIs. Panel 2: peer agreement carries most of the ranking, while target commitment adds 1.1–10.1 AUROC points; the gray diamond is each dataset's strongest external encoder and chance (.5) is marked. Panel 3: on the three 512-question panels, the ranking gain becomes maximum coverage under a 5% observed accepted-pool error ceiling, read from the full-label risk–coverage curve. All methods use the same fixed answers, source labels, and evaluation rows.

Peer readout (AUROC)	GSM8K	TriviaQA	MMLU-Pro	GPQA	MedQA	TruthfulQA
C: agreement with the four peers' own answers	.837	.922	.835	.704	.866	.646
J: the same four peers judge the frozen answer	.993	.908	.881	.845	.929	.785
A+C	.938	.933	.859	.763	.892	.764
A+J	.994	.910	.882	.846	.928	**.799**
A+C+J	**.996**	**.938**	**.884**	**.851**	**.929**	**.799**

Table 2: **The same peers, read as a verdict, carry more than their agreement count on five of six datasets.** Same frozen answers, double holdout, and head as Table 1, on the rows where every peer returned a verdict; the full ladder, exclusions, proper scores, and call costs: Appendix G.

Table 1 reports the comparison. Pooled over environments, peer corroboration alone outperforms every text predictor, and the minimal A+C fusion beats every standalone view and comparator on both displayed metrics wherever they are measured. Adding eight target resamples gives the best measured configuration on the three 512-question panels, with its clearest gain on GSM8K. On those panels, one target–peer agreement statistic already beats every text predictor (Appendix N). One exception to the pooled margins is systematic. On the strongest target of every 512-question panel, Sonnet 4.5, a single self-report already ranks above the fusion (Appendix H), the case Section 4.3 explains.

The two comparators locate the boundary. The perturbation composite wins only on GSM8K, where the target's own behavior is most informative. There it beats the peer view alone and, paired with self-report, edges the fusion. On TriviaQA and MMLU-Pro both comparisons reverse (Appendix G). The released correctness model ranks above every retrained encoder and its mean score tracks each target's accuracy across models, yet it ranks below the peer view alone at every pooled dataset point estimate. The retrained encoders keep useful ordering within each target model's environment and lose level alignment across models (Appendix H), exactly the environment-local component the next section isolates.

Table 2 reads the same peers two ways. Asking them to judge the frozen answer beats counting their agreement on five of six datasets and is level on TriviaQA, and the gain on GPQA, where the panel is right only about half the time, matches GSM8K's, so a useful verdict does not require a panel that is usually right. A weaker external verifier loses to agreement on four of six datasets, so what a verdict adds still follows the judge's capability (Appendix G). The fused head takes both readouts and matches or improves the better two-view fusion in every column.

The same fixed-answer protocol extends to GPQA-Diamond (Rein et al., 2024), MedQA (Jin et al., 2021), and TruthfulQA (Lin et al., 2022b) (Table 3), and it survives a change of domain. Holding out the dataset and the model at once, the ranking exceeds chance in all fifteen environments while the probability level moves (Appendix K).

4 Monotone links preserve the ranking; separation predicts its limits

The transfer result separates two questions. Why can the same ordering survive a new model and new questions, and when does peer agreement contain correctness information at all? The first is an invariance statement, which we call *order invariance*; the second an information model, whose central fact is a *label-free identifiability* limit.

4.1 Ranking transfers while the probability scale moves

An *environment* e is a target answer model and dataset. Let S^* denote the source-trained A+C score, frozen before the target environment is evaluated.

Assumption 1 (monotone link invariance)**.** *For every environment e, the score-conditional correctness probability $g_e(s) = \mathbb{P}_e(Y = 1 \mid S^* = s)$ is strictly increasing on the support of S^*.*

A new environment may stretch or shift the score–probability relationship without reversing its order. Thresholding S^* is equivalent to thresholding its score-conditional posterior, so within an environment the two have identical ROC and risk–coverage curves (Proposition 1), and decisions

use only that ordering. The statement is conditional. It says what a monotone link buys within one environment and is silent on why source training produces one. Under an affine refinement of the link (Working Model 1), pairwise comparisons also cancel the environment offset b_e. A small labeled audit diagnoses the link and, where the affine form holds, repairs the environment-local scale (Appendices M, I, K).

4.2 Unanimity is not truth

Agreement is evidence only when peers are more likely to join a right answer than a wrong one, and unanimity is where shared misconceptions can break that logic. A regular agreement model (Dawid & Skene, 1979) has one decisive pair of rates: a peer joins the target's answer with probability π when the target is right and c when it is wrong. But real panels also share misconceptions and can converge on the same wrong answer. On such a question, agreement measures the popularity of a belief, which can diverge from truth. The consensus-attractor model (Definition 1) therefore adds a question-level component to the regular rates: with probability α a question is an attractor on which all $m + 1$ panel members share one cluster that is correct with probability τ; otherwise the target is correct with probability ρ and each of m peers joins it with probability π or c.

Below unanimity, the attractor component is absent and the posterior log-odds is affine in the agreement count k, with slope $L = \log\{\pi(1-c)/(c(1-\pi))\}$. Inside the model the count is the likelihood-ratio statistic, so it is sufficient for the full peer-join vector (Appendix M). At unanimity, attractor mass enters directly. As the panel grows, unanimity from the regular component vanishes geometrically while the attractor mass is unchanged, so when attractor questions exist ($\alpha > 0$) the purity of unanimous agreement converges to the attractor rate τ and not to one (Proposition 4). On SimpleQA's eight-model panel unanimity purity rises only from about one half to two thirds as the panel grows to seven members (Appendices G and N). More peers buy more evidence on regular questions and cannot certify safety at unanimity. Within this agreement law, whose attractors are perfectly unanimous, τ enters the joint law only through correctness itself. The unlabeled agreement count therefore cannot distinguish a right consensus from a wrong one: its distribution does not depend on τ, every test built on unlabeled counts has power equal to its size, and labeled anchors from the unanimity cell estimate the cell's purity at the parametric rate (Proposition 2; Appendix M). Labels meant to estimate unanimity correctness must come from unanimous questions. Such labels estimate the cell's purity $\mathbb{P}(Y{=}1 \mid k{=}m)$ itself, which equals τ only in the large-panel limit, and they do not identify which unanimous answers are wrong. The model was validated before it was used to predict. It beats its no-attractor ablation out of sample, and Proposition 5 explains why that ablation is structurally misspecified. A confirmatory pass preregistered before the 512-question panels existed recovered five of six predictions, and a preregistered ablation confirms that the agreement fraction carries nearly all of the agreement view (Appendix N).

4.3 Separation predicts the boundary

Away from unanimity, a panel's value is set by how well its members separate right target answers from wrong ones. On regular questions, panel information is exactly that separation between the two Bernoulli join laws. Writing J for their symmetrized Kullback–Leibler divergence,

$$(\pi - c)L = J(\pi, c). \tag{2}$$

A panel need not be accurate on average; it must join a right target answer more often than a wrong one. The identity makes the coming failure predictable, since panel value reduces to two join rates and anything that pushes π toward c silently empties the agreement signal. Larger fitted separation means better agreement-based ranking across the thirty source environments (Figure 3; Appendix N). The attractor parameters (α, τ) set the independent unanimity boundary.

The source generation exposed one failure mode: when the target was much stronger than its peers, π approached c and agreement lost its ranking signal. On the strongest target of each 512-question panel this is already visible (Section 3). Source data alone cannot tell how a new target stands relative to its panel, and four source-only probes fail at exactly that term (Appendix N). So the estimator keeps self-report beside peer agreement, and a new environment still gets a small link audit. The blind spot predicts what a model-generation change should do: stale agreement should hurt once a target outruns its panel; refreshed peers should restore separation; mixing stale peers back in should dilute it.

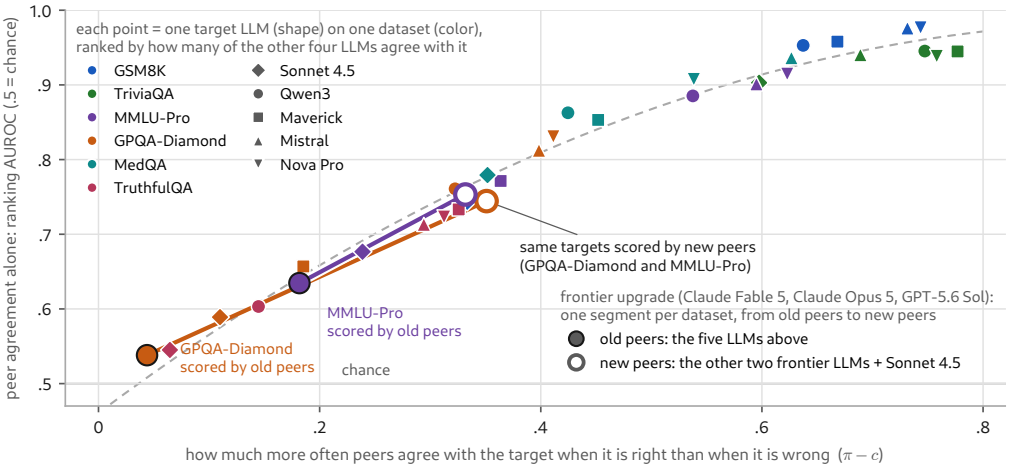

Figure 3: **Peer agreement is worth about as much as the panel separates right from wrong target answers.** Each point is one target model on one dataset (thirty source environments); the horizontal axis is the labeled join-rate gap $\pi - c$ estimated on that environment's rows, the vertical axis the ranking AUROC of the agreement count alone. The relationship is observed and monotone; the gap alone does not fix the AUROC (Appendix M). The frontier upgrade of Section 5 moves along the same curve: scored by the old peers, the frontier targets sit at the left end; scored by refreshed peers, they move back up.

Evaluation setting	Panel	n	A (self-report)	C (agreement)	A+C
GPQA-Diamond	held-out	984	.719	.704	**.770**
MedQA	held-out	2,560	.765	.867	**.892**
TruthfulQA	held-out	2,551	.749	.646	**.764**
Frontier GPQA	stale	567–572	.784	.522	.664
	refreshed	527–531		.737	.772
	refreshed + stale	572		.620	–
Frontier MMLU-Pro	stale	1,530–1,532	.691	.622	.683
	refreshed	1,522–1,524		.739	**.778**
	refreshed + stale	1,522–1,524		.689	.735

Table 3: **The fusion extends across new datasets; stale peer agreement fails at a generation gap.** The upper block applies the same held-out protocol to GPQA-Diamond, MedQA, and TruthfulQA. The lower blocks test the separation prediction: stale, refreshed, and mixed panels under a previous-generation calibrator with zero target labels on two datasets; refresh restores agreement, while mixing stale peers dilutes it. Self-report does not depend on the panel, so its value appears once per dataset. Details: Appendices L and O.

5 A MODEL UPGRADE CAN BREAK STALE CORROBORATION

This section tests the prediction of Section 4.3 where it should hurt most, twice: three frontier models answered GPQA-Diamond, and later MMLU-Pro, while a previous-generation calibrator and panel scored them. On GPQA the targets outperformed the old panel, whose accuracy averaged about one half, by a wide margin; on MMLU-Pro the gap was moderate. The account has two dials: how much the peers know relative to the target decides how much of their evidence survives, and how they are read decides how much of it is extracted; the target's own report depends on neither.

The upper block of Table 3 extends the fusion to three new datasets. The lower blocks are the generation-gap tests: GPQA is the discovery instance under a frozen estimand, MMLU-Pro its preregistered replication, five of five. Stale agreement loses about half of its ranking signal at the moderate gap and all of it at the wide one, pulling the fusion below self-report. Replacing the panel restores agreement, on GPQA from .522 to .737, and pooling stale peers back in dilutes it, because near-independent peers average their noise into the pooled fraction (a head trained with frontier labels that keeps the two counts separate is not diluted; Appendix O). At the wide GPQA

Scorer at the frontier (AUROC, common rows)	GPQA-Diamond ($n = 453$)	MMLU-Pro ($n = 1,431$)
A (target self-report)	.789	.668
C, stale panel (agreement)	.483	.593
J, stale panel (three previous-generation peers judge)	.641	.661
Qwen3-32B judge	.570	.584
Released correctness model	.495	.499
Calib-text encoder (retrained)	.459	.449
C, refreshed panel (agreement)	.661	.701
J, refreshed panel (judge)	.778	**.786**
A+C, refreshed panel	.730	.753

Table 4: **Every fixed previous-generation scorer degrades at the generation gap, the same peers asked to judge included; refreshed judges recover to self-report at the wide gap (no detected difference) and above it at the moderate one.** Rows on which every scorer is valid, one denominator per column; full populations, per-judge values, sensitivities, and paired intervals: Appendix O.

gap, the refreshed fusion is indistinguishable from self-report on this sample, so the deployable choice there is self-report plus a labeled audit. Refreshed peers restore an option without beating it. At the moderate MMLU-Pro gap, the refreshed fusion beats self-report, and the stale calibrator's probabilities sit far below the frontier targets' accuracy until a twenty-label intercept refit closes most of the gap.

Agreement is not the only casualty. On the rows where every scorer is valid (Table 4), the same three previous-generation peers asked to judge lose about .2 AUROC per judge on the same questions once the target is a frontier model. The single judge, the correctness model, and the retrained encoder also lose ranking power, none reaches the refreshed panel, and self-report holds. Refreshed judges recover, above self-report at the moderate gap and indistinguishable from it at the wide one. Inside the refreshed panel the one previous-generation member is the weakest judge and dilutes the pooled verdict (Appendix O). The separation account predicted the agreement half of this; the verdict half was measured under a frozen plan, and the remaining scorers are consistent with it.

6 Decisions and information boundaries

The transferred score supports four deployment contracts, in increasing order of label cost: fixed-budget routing, arbitration between two models, automation under a stated error ceiling, and certified automation. Its advantage over a text predictor lies in adaptation cost, adding model calls at inference but no model-specific training. Counting agreement costs nothing beyond the peers' answers, plus one clustering call on open answers; a verdict costs one call per peer and reads more (Section 3). Both weaken across a generation change without a refreshed panel; choosing between them is a labeled decision we do not automate. Fixed-budget routing verifies the least confident answers. Under Assumption 1 it uses the transferred ordering alone, so beyond the first-deployment link audit it needs no target labels.

Arbitration between disagreeing models asks for more, one probability scale both models' scores can share. On it, comparing the two answers improves the choice on average (Appendix P).

A stated error ceiling needs target-local threshold calibration. The source-selected scale transfers unreliably, so the threshold requires an in-environment labeled audit, and choosing among the raw, transferred, and repaired scales is itself a labeled decision. Small audits choose badly where the scale is already calibrated, and the selection becomes safe only near a hundred labels (Appendices J and K). As a descriptive full-label operating point, A+C accepts 92.5% of TriviaQA answers at 5% observed error, against 35% for the best retrained encoder.

A guarantee costs more labels than an audit. With roughly a hundred labeled anchors, an exact-binomial threshold walk certifies a distribution-free risk bound and at 10% risk recovers most of the oracle coverage on TriviaQA (Appendix Q).

One boundary admits no repair from the measured self- or peer-agreement signals. When target and peers share a stable misconception, self-consistency and unanimity can be high on a wrong answer. On the thirty TruthfulQA questions where all five models give the same wrong answer, resample self-agreement averages .861 against .712 on other errors, so the answers models share wrongly

are those they hold most stably. Reducing it requires capability-heterogeneous or external evidence (Appendix L).

7 Related work

Surveys organize black-box confidence into self-report, consistency, and surrogate models (Geng et al., 2024). Views A and B build on verbal self-evaluation (Lin et al., 2022a; Kadavath et al., 2022; Tian et al., 2023) and sampling-based uncertainty (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024). Three lineages are closest.

First, invariance as an organizing principle: Kuhn et al. (2023) built uncertainty on a linguistic invariance, treating answers with one meaning as one outcome. Ours is statistical, asking which score properties survive an environment change.

Second, panels of models: peer signals appear in prior work as surrogate scoring, cross-model checks, encoder features, and consistency weights (Shrivastava et al., 2023; Zhang et al., 2023; Xia et al., 2025; Ning et al., 2025), Ding (2026) audits majority-answer agreement, Xue et al. (2026) call one verifier when the target's self-consistency is uncertain, and Hamidieh et al. (2026) weight a scale-matched ensemble uniformly, the assumption Section 5 breaks. These works use other models as evidence. We model when that evidence is real, read the same peers as a count and as a verdict on one fixed answer, and evaluate on models and questions the supervision never saw.

Third, transfer as the evaluation: the closest protocol, Pedapati et al. (2025), reports ranking under model shift. We study calibrated probabilities under proper scores, which component transfers, and its failure boundary. Xiao et al. (2025) pursue the same transfer with a text predictor trained on many models' correctness histories. Their released model is our strongest text comparator and ranks below peer agreement alone at every pooled dataset point estimate. APRICOT (Ulmer et al., 2024) and Calib-n (Xia et al., 2025), the closest supervised predecessors, are retrained inside our source splits from their original checkpoints with their published hyperparameters, source-only validation selection, one seed, and, for Calib-n, the unseen-model adaptation of Appendix F.

Risk-control methods (Bates et al., 2021; Jung et al., 2025) combine a score with an in-environment labeled sample to place a certified acceptance threshold. Section 6 states when that local audit is required. Full positioning, including Thermometer (Shen et al., 2024): Appendix R.

8 Conclusion

Freezing the answer reveals two dominant signals: self-report and sampling mostly measure commitment; peer evidence adds information the target cannot supply. The fused head transfers across held-out models and questions and beats retrained encoders on four panels and, on the three 512-question panels, a released correctness model and a five-family perturbation composite. Asked to judge, the same peers give more than when counted on five of six datasets.

The account behind the result predicted its own failure. A monotone environment link transfers the ranking and leaves the probability scale local, so ranking costs no target labels where the link holds and the scale needs a local audit. Peer agreement is governed by the panel's separation between right and wrong target answers. In the observed upgrades separation weakened, and agreement with it, to chance at the wide gap. The same peers weakened as judges, with every other fixed scorer. Refreshed peers restored both readouts, above self-report at the moderate gap and indistinguishable from it at the wide one, while self-report held. Within the agreement model, unanimity correctness leaves the unlabeled agreement law unchanged, so estimating it requires labels from unanimous questions.

The evidence covers short-answer question answering, one generation change, and four datasets with exact encoder retraining (Appendix A). The protocol needs only an answer that can be frozen and graded, which points to long-form generation, where the frozen object is a claim, and to a labeled rule for choosing the peer readout. What transfers is a working rule: freeze the answer, separate commitment from corroboration, read peers as judges and refresh them when the target moves, audit the environment-local link with labels, and never treat unanimity as truth.

AI USE STATEMENT

The research questions, methodology, experimental decisions, and scientific conclusions of this work are the authors'. Generative AI tools assisted under the authors' direction: at every step, from the research idea to the design of each experiment and the reading of its results, the tools proposed and the authors decided. In this work, we used generative AI tools for developing the theoretical framework, mathematical claims, and proofs; proposing and refining hypotheses; designing experiments and giving feedback on methodology; implementing data collection and analysis; cleaning and reformatting data; and interpreting results. We have not used generative AI tools to generate synthetic evaluation data; translation and qualitative or thematic analysis are not applicable to this work. Additionally, we used generative AI tools for drafting and editing text and code, figure design and production, literature searches, and reference formatting. We have reviewed all AI-assisted work: every design decision, experiment, and claim was decided by the authors after discussion with the tools; AI-proposed hypotheses were treated as hypotheses, preregistered before their data were collected, and several were refuted, as reported; proofs were checked step by step and against numerical tests; every reported number was traced to an archived artifact by automated checks; every citation was verified against the original source. We did not use AI to fabricate, impute, or select results. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The evaluation protocol, feature definitions, and estimator are specified in Section 2.3; data construction and freezing rules in Section 2.1; complete proofs in Appendix M; and protocol safeguards (gold isolation, bounded confidence-line parsing, strict final-block sensitivity, truncation-robustness refits) in Appendix D, together with the analysis plans fixed before their data were collected. Dataset identities are stated in the paper, and the appendices describe each experiment's models, instruments, and configurations.

REFERENCES

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535.

Dylan Bouchard and Mohit Singh Chauhan. Uncertainty quantification for language models: A suite of black-box, white-box, LLM judge, and ensemble scorers. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=WOFspd4lq5.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Kaihua Ding. When LLMs agree, are they right? auditing self-consistency and cross-model agreement as confidence signals. *arXiv preprint arXiv:2607.08065*, 2026.

Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 973–978, 2001. URL https://cseweb.ucsd.edu/~elkan/rescale.pdf.

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.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. 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. 6577–6595, 2024. doi: 10.18653/v1/2024.naacl-long.366.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Kimia Hamidieh, Veronika Thost, Walter Gerych, Mikhail Yurochkin, and Marzyeh Ghassemi. Complementing self-consistency with cross-model disagreement for uncertainty quantification. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lOoRJo8xWy.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017. doi: 10.18653/v1/P17-1147.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UHPnqSTBPO.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Dharshan Kumaran, Arthur Conmy, Federico Barbero, Simon Osindero, Viorica Patraucean, and Petar Veličković. How do LLMs compute verbal confidence? *arXiv preprint arXiv:2603.17839*, 2026.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in LLMs: Reward calibration in RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l0tg0jzsdL.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022a. ISSN 2835-8856. URL https://openreview.net/forum?id=8s8K2UZGTZ.

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 (Volume 1: Long Papers)*, pp. 3214–3252, 2022b. doi: 10.18653/v1/2022.acl-long.229.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=DWkJCSxKU5.

Xin Liu, Muhammad Khalifa, and Lu Wang. LitCab: Lightweight language model calibration over short- and long-form responses. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jH67LHVOIO.

Potsawee Manakul, Adian Liusie, and Mark 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*, pp. 9004–9017, 2023. doi: 10.18653/v1/2023.emnlp-main.557.

Putra Manggala, Atalanti A. Mastakouri, Elke Kirschbaum, Shiva Kasiviswanathan, and Aaditya Ramdas. QA-calibration of language model confidence scores. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=D2hhkU5O48.

Kun-Peng Ning, Shuo Yang, Yuyang Liu, Jia-Yu Yao, Zhenhui Liu, Yonghong Tian, Yibing Song, and Li Yuan. PiCO: Peer review in LLMs based on consistency optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sfQ6XpApfS.

Tejaswini Pedapati, Amit Dhurandhar, Soumya Ghosh, Soham Dan, and Prasanna Sattigeri. Large language model confidence estimation via black-box access. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=WrWYChkyRI.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002. doi: 10.1162/089976602753284446.

Jianru Shen. Provable limits and certified deferral for verbalized uncertainty in small language models. *arXiv preprint arXiv:2608.05064*, 2026.

Maohao Shen, Subhro Das, Kristjan Greenewald, Prasanna Sattigeri, Gregory W. Wornell, and Soumya Ghosh. Thermometer: Towards universal calibration for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44687–44711. PMLR, 2024. URL https://proceedings.mlr.press/v235/shen24c.html.

Vaishnavi Shrivastava, Percy Liang, and Ananya Kumar. Llamas know what GPTs don't show: Surrogate models for confidence estimation. *arXiv preprint arXiv:2311.08877*, 2023.

Ao Sun, Tian Sun, and Jiaxing Geng. The score granularity gap in black-box LLM classification: A comparative study of confidence constructions. *arXiv preprint arXiv:2606.22179*, 2026.

Hexiang Tan, Fei Sun, Sha Liu, Du Su, Qi Cao, Xin Chen, Jingang Wang, Xunliang Cai, Yuanzhuo Wang, Huawei Shen, and Xueqi Cheng. Too consistent to detect: A study of self-consistent errors in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4755–4765, 2025. doi: 10.18653/v1/2025.emnlp-main.238.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023. doi: 10.18653/v1/2023.emnlp-main.330.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Joon Oh. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15440–15459, 2024. doi: 10.18653/v1/2024.acl-long.824.

Jiawei Wang, Yanfei Zhou, Siddartha Devic, and Deqing Fu. Are LLM decisions faithful to verbal confidence? *arXiv preprint arXiv:2601.07767*, 2026.

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

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Sean Wu, Fredrik K. Gustafsson, Edward Phillips, Boyan Gao, Anshul Thakur, and David A. Clifton. BAS: A decision-theoretic approach to evaluating large language model confidence. *arXiv preprint arXiv:2604.03216*, 2026.

Yuxi Xia, Pedro Henrique Luz De Araujo, Klim Zaporojets, and Benjamin Roth. Influences on LLM calibration: A study of response agreement, loss functions, and prompt styles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3740–3761, 2025. doi: 10.18653/v1/2025.acl-long.188.

Hanqi Xiao, Vaidehi Patil, Hyunji Lee, Elias Stengel-Eskin, and Mohit Bansal. Generalized correctness models: Learning calibrated and model-agnostic correctness predictors from historical patterns. *arXiv preprint arXiv:2509.24988*, 2025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Yihao Xue, Kristjan Greenewald, Youssef Mroueh, and Baharan Mirzasoleiman. Verify when uncertain: Beyond self-consistency in black box hallucination detection. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=6tlLISSgiu.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. SAC^3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15445–15458, 2023. doi: 10.18653/v1/2023.findings-emnlp.1032.

A THE FAILURE MAP

Table 5 maps each claim of the paper to the test that could have refuted it; Table 6 then collects every boundary documented in the study, where each is shown, and the tested repair or operating guidance.

Claim	Decisive test	Evidence	Verdict	Where
Target signals measure commitment	shared direction; all-wrong questions	observational	one shared direction; stable errors are the most self-consistent	§2.2, App. C
Fusion beats retrained encoders	double holdout, exact retraining, four panels	observational	wins every dataset and metric	Tab. 1, App. H
Fusion beats a released correctness model	zero-shot 8B model on the same rows	preregistered	headline PASS on 3/3 datasets; 3 of 4 auxiliary predictions PASS	App. G
Fusion beats the perturbation composite	five families, same holdout	preregistered	headline PASS on 3/3 datasets; 3 auxiliary predictions FAIL, all on GSM8K	App. G
One count carries the agreement view	count-only ablation	preregistered	PASS; shape statistics add little	App. N
Agreement matches the peers' own verdict	same peers judge the frozen answer, same rows and head	preregistered	not met: verdict above count on GSM8K, MMLU-Pro, GPQA, MedQA, and TruthfulQA, level on TriviaQA; the head taking both is never behind	Tab. 2, App. G
Fixed source-side scorers hold at the frontier	correctness model, retrained encoder, two single judges, and the same three previous-generation peers asked to judge, on the frontier targets' answers	preregistered	not met: all degrade below self-report; per-judge paired loss .21–.23 on MMLU-Pro (3 of 3 intervals exclude zero), .11–.23 on GPQA (2 of 3); stale verdicts stay above stale agreement (registered within-.05 clause FAILED)	Tab. 4, App. O
Refreshed judges beat stale judges, refreshed agreement, and self-report (registered)	the Table 3 refreshed panel asked to judge the frontier answers	preregistered	refreshed above stale judges on MMLU-Pro (interval excludes zero), GPQA interval includes zero (20 wrong answers); refreshed judge at or above refreshed agreement PASS; above self-report on GPQA FAILED (no detected difference)	Tab. 4, App. O
Strongest single peer matches the four-peer count	one-peer agreement bit vs. fraction	preregistered	not met: below on GSM8K and TriviaQA, above on MMLU-Pro	App. G
Peer construction matches or beats TU with self-report held fixed	A+TU vs A+C, same rows and folds; registered A+C $>$ TU in at least 14 of 15 environments	observational + preregistered	A+C ahead on 2/3; level on TriviaQA; the 14-of-15 environment prediction FAILED (12 of 15); joint TU partition is not strict LOMO	App. R
Ranking transfers, scale stays local	cross-dataset holdout; link diagnostics	observational	above chance 15/15; scale moves	App. K, I
Unanimity correctness needs labels	Prop. 2; purity vs. panel size	theorem + observational	purity tends to τ rather than to one	App. M, N
Stale panel collapses; refresh restores; pooling dilutes	frontier GPQA, three targets	estimand preregistered; direction exploratory	collapse, refresh, dilution observed; registered clause "fusion still beats self-report" FAILED (fusion back to self-report)	Tab. 3, App. O
Same at a moderate gap	frontier MMLU-Pro, three targets	preregistered replication	5 of 5 PASS; fusion above self-report; registered provider-disjoint sensitivity: collapse unchanged, refresh gain smaller	Tab. 3, App. O
Small audits choose repairs badly	selection at 10–120 labels	preregistered	1 of 4 PASS; safe near 100 labels	App. K
Fusion holds at equal audit budgets	five scores, same labels, own best repair	observational	lowest NLL in 9–10 of 15; never behind raw or the correctness model	App. K

Table 5: **Each claim and the test that could have refuted it.** "Evidence" states whether the test was observational, preregistered before its data existed, or a proved statement checked against data. Failed registered predictions are kept as failures and read as boundaries, here for the headline claims and in Appendices G, K, N, and R for the secondary registered predictions; Table 6 collects the documented boundaries with their consequences.

Full limitations. The frontier upgrade experiment covers one generation change, three targets, and two datasets. GPQA-Diamond and MMLU-Pro are public, so frontier-model training exposure cannot be ruled out. The experiment establishes behavior under the observed target–panel gap rather than attributing that gap uniquely to general capability. The panel-refresh result uses a frontier-dominant panel rather than a purely same-generation one, and the frontier arbitration count

Failure mode	Where shown	Repair or guidance
Probability scale is environment-local	Sec. 4.1; Apps. I–K, O	under the monotone link, ranking and fixed-budget routing are unaffected; when absolute probabilities are needed, audit raw, transferred, and repaired scales; if repair is selected, $\sim$20 anchors estimate it (93–97.5% of the oracle gap here; heterogeneous across domains, can hurt near-identity links); choosing whether to repair needs 80–120 labels unless the shift is large (App. K)
Shared misconceptions defeat agreement, self and peer	Sec. 6	no repair from the measured self/peer signals; needs capability-heterogeneous or external evidence
Stale panel collapses at a generation gap, counted or asked to judge	Sec. 5	refresh restores both readouts (and adds a gain over self-report at a moderate gap; the refreshed verdict is at or above refreshed agreement); compare with dropping the peer view; do not pool old peers into one agreement fraction, keep them as a separate feature or drop them
No same-generation peer is available	Sec. 5	use the transferred self-report ranking (A alone, .784 at the frontier gap) plus a small labeled audit; treat the peer view as unavailable rather than stale
Double shift (new model and new dataset) degrades stale-panel fusion	App. O	trust the transferred self-report ranking; drop the stale panel
Cheap peers are dataset-dependent	App. G, L	mixed on standard benchmarks (better on GSM8K, worse on TriviaQA and MMLU-Pro); in the hard regime the agreement view degrades (interval excludes zero) and the fusion contrast is inconclusive
Affine link insufficient	1 of 15 environments (App. I)	isotonic recalibration in that cell
Raw confidence beats calibrated arbitration	GSM8K (App. P)	known heterogeneity; calibrated wins pooled and on four of five datasets
Shared acceptance threshold transfers poorly	App. J	per-environment threshold, set with labels
Strict source-selected ceiling: fusion behind C-only	App. J	if a source threshold must transfer untouched, C-only meets the requested risk in more environments (9 of 10 on GSM8K and TriviaQA against 5 of 10 for A+C); with an in-environment audit, place the threshold directly rather than choose the score by source-threshold transfer
Frontier model returns no visible answer	App. O (Fable 5: 20 of 198; 22 of 594 overall)	count as incorrect on the request denominator; route to review

Table 6: **Different failures require different repairs.** The map records boundaries observed across the study, where each is documented, and the tested repair or operating guidance. The one failure with no repair from the measured self/peer signals, shared misconceptions, is an information boundary: agreement detects instability, and a stable collective belief is what it cannot detect.

is eighteen pairs, reported as an illustration. Anchor-repair percentages are measured at the generation gap; the shared frontier intercept pools labels across three target-model environments because their drifts align, an empirical economy rather than a guarantee of Proposition 3. Cross-domain anchor value is heterogeneous, and on near-identity links small anchor sets can hurt.

GPQA and MedQA extend the fusion result but not the encoder comparison, which remains scoped to the four datasets where the baselines were retrained. SimpleQA remains a 128-question panel with unanimity cells of three to seven questions. Source-only attractor parameters predict dataset-level link slopes and fail on target-level deviations. The model compresses dataset-level link behavior; target-specific forecasting requires target-local evidence. No universal label budget is established for diagnosing a stale panel. The attractor component is question-level, but its cause is not identified: intrinsic ambiguity and correlated errors from common training exposures are confounded, so provider diversity is a proxy rather than an independence guarantee. The affine link is a working approximation, flagged in three of fifteen environments and materially beaten by isotonic recalibration in one. The transfer result repeats across five held-out target models per 512-question panel, which does not estimate uncertainty over a broader population of future model families. The behavioral pilot has one target model. Semantic clustering and correctness grading can introduce errors; TriviaQA labels required explicit adjudication. The cross-provider judge evidence covers one panel, TriviaQA. The peers' verdict at the generation gap was measured for one previous-generation and one refreshed panel at a single registered reasoning budget; a separate 16,000-token attempt and a fail-closed treatment move no contrast (Appendix O). The peer comparison is scoped to the released, paper-aligned training budgets, and the text comparators span 110M-parameter encoders to the 8B correctness model; the paper does not test whether larger text predictors close the remaining gap to peer agreement.

We do not claim an inference-cost advantage: deployable A+C adds six model calls, well under a cent per answer and dominated by the target's own self-report rather than by the peers (Appendix F). Our advantage is on the adaptation side, where the encoder recipes require GPU fine-tuning on thousands of in-domain labels from the deployed models. The ranking estimator itself uses no target labels under the monotone-link assumption; validating that link or an absolute probability may still need the small target audit described above.

All panels are short-answer question answering, chosen because a fixed answer there has a checkable equivalence class; long-form generation would require scoring claims within an answer (Manakul et al., 2023) and is future work. Finally, every calibrated probability is tied to a deployment distribution: link parameters are indexed by the model–dataset pair jointly, so anchors from one dataset do not transfer to another.

B What existing methods observe

Verbal self-evaluation. Direct confidence, P(True), top-k probability allocation, and multi-step self-probing ask the model to assess its answer (Kadavath et al., 2022; Tian et al., 2023; Xiong et al., 2024). They are cheap, but the report can depend on wording, examples, score visibility, and response format.

Repeated sampling. Self-consistency, SelfCheckGPT, semantic entropy, and graph-based methods measure how concentrated repeated generations are (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024; Lin et al., 2024). They avoid asking for a confidence number, but a model can repeatedly produce the same misconception. These methods score the spread of generations, or, as SelfCheckGPT does, the sentences of one generation against sampled passages; our comparison evaluates their induced support and entropy as confidence for the frozen answer.

Perturbations. Another approach changes the presentation of the input and measures whether the answer is stable. Pedapati et al. use stochastic decoding, paraphrases, sentence permutations, entity repetition, stopword removal, and response consistency as features in a supervised logistic predictor (Pedapati et al., 2025). This is a strong baseline for our setting. It also exposes a control requirement: a perturbation is valid only if it preserves the question being answered.

Surrogate and peer models. An open model's probability can serve as a surrogate for a closed model, and disagreement across models can detect some failures hidden by self-consistency (Shrivastava et al., 2023; Hamidieh et al., 2026). These methods fail when models share the same misconception, so consensus is evidence rather than truth. Under our held-out protocol, the total-uncertainty estimator of Hamidieh et al. (2026) ranks below the fusion in twelve of fifteen environments and has the worse proper score in fourteen (Appendix R).

Independent evidence. Retrieval, a task-specific verifier, or an independent critic can directly test the answer. Unlike self-report and sampling, these methods add information about correctness. We treat them as a separate evidence class rather than treating every signal as interchangeable uncertainty.

C Pilot details

Formal view definitions. Stage 1 generates and freezes the object:

$$a_i \sim M(\cdot \mid q_i), \qquad o_i = (q_i, a_i), \qquad y_i \in \{0, 1\}. \tag{3}$$

For n fixed objects and d_A self-report instruments, the self-report matrix is

$$X^{(A)} \in [0,1]^{n \times d_A}, \qquad X^{(A)}_{ik} = \text{reported probability that fixed answer } a_i \text{ is correct.} \tag{4}$$

Here an *instrument* is one fixed elicitation prompt that returns a probability that the frozen answer is correct; one instrument gives one column.

The behavioral view is defined per temperature by

$$S_i(T) = \frac{1}{5}\sum_{r=1}^{5} \mathbb{I}\{a_{ir}(T) \simeq a_i\}, \qquad H_i(T) = -\sum_c p_{ic}(T)\log p_{ic}(T), \tag{5}$$

where $a_{ir}(T)$ is the r-th rerun answer, $\simeq$ is semantic match to the fixed answer, and $p_{ic}(T)$ are the rerun cluster frequencies. Stacking the behavioral rows $x_i^{(B)} = [S_i(.2), S_i(.7), S_i(1), H_i(.2), H_i(.7), H_i(1)]$ gives $X^{(B)} \in \mathbb{R}^{n\times d_B}$. The four peer-agreement statistics of View C form

$$X^{(C)} \in \mathbb{R}^{n\times 4}. \tag{6}$$

The full supervised dataset is

$$X = [X^{(A)} \mid X^{(B)} \mid X^{(C)}] \in \mathbb{R}^{n\times d}, \qquad \boldsymbol{y} = [y_1, \ldots, y_n]^\top. \tag{7}$$

Table 7 summarizes the data objects of Section 2.1 and the failure mode characteristic of each view.

Data object	Shape	Source	Main failure mode
$X^{(A)}$	$n \times d_A$	verbal reports	prompt reactivity
$X^{(B)}$	$n \times d_B$	repeated behavior	stable misconception
$X^{(C)}$	$n \times d_C$	peer answers	correlated peer error
$\boldsymbol{y}$	$n \times 1$	independent grading	label or grader error

Table 7: **The fixed-answer control makes every signal refer to one object.** The rows separate the three measured views, independent grading, and each one's characteristic failure mode.

The observed CCA canonical correlations between the six self-report and six behavioral columns are $(.460, .281, .180, .119, .049, .006)$; only the first exceeds its paired-permutation 95% null threshold ($p = .010$; second component $p = .162$). With 114 objects and moderate behavioral reliability we describe this as one detected shared direction. It does not establish that the true shared dimension is exactly one.

Behavioral support for the fixed answer is non-increasing in temperature for 111 of 128 questions; the intermediate temperature gives mean support .781. Split-half Spearman reliability of the support statistic is .637, .592, and .590 at $T = .2, .7, 1$. Table 8 gives the pilot prediction comparison summarized in Section 2.2.

Features	AUROC	Brier score
Self-report A	$.801 \pm .014$	.145
Behavior B	$.723 \pm .013$	.158
A+B	$.809 \pm .018$	.143

Table 8: **Repeated behavior adds little beyond self-report in the pilot.** AUROC is mean $\pm$ standard deviation over 30 repeated five-fold evaluations; Brier score is the corresponding mean.

Main-text pilot numbers, in full. Complete pairing across all six self-report columns leaves 114 objects, of which 27 are correct. Mean support for the fixed answer falls from .913 at $T = .2$ to .703 at $T = 1$; low-temperature support is nearly saturated (correctness AUROC .558), while support at $T = 1$ reaches .736 and negative semantic entropy .755, because higher temperature reveals competing answers. The commitment direction's largest-magnitude loading is semantic entropy at $T = 1$ ($-.967$). Under repeated five-fold evaluation, self-report alone reaches AUROC $.801 \pm .014$, behavior alone $.723 \pm .013$, and the combination $.809 \pm .018$; the combined-minus-self-report AUROC difference carries a question-bootstrap 95% interval of $[-.064, +.069]$.

Self-report instruments. All six pilot instruments share one evaluation frame and differ only in what the prompt displays: direct asks for the probability alone; scored-0 and scored-100 show a submit-or-abstain scoring rule with no penalty and a -100 penalty for a wrong submission; threshold adds an explicit submit-only-above-a-stated-confidence rule; salience asks the model to restate the scoring rule before answering; expected-value asks for a step-by-step expected-score comparison before deciding.

D PROTOCOL SAFEGUARDS

Invalid behavioral run. An initial 200-call run used a system prompt that differed slightly from the Stage-1 prompt and changed behavior materially. It is archived separately and is not pooled with the primary 1,920-call run.

Paraphrases. Paraphrase features require an independent equivalence check because ordinary rewrites can change the question's scope and requested answer type; a pilot probe under a three-validator unanimity gate found no incremental correctness value, and a preregistered full-scale run confirms it (Appendix G).

Missing reports. Instrument parse failures are recorded as missing. They are not filled with an arbitrary confidence such as .5, which would mix the observation process with the confidence value.

Gold isolation. Gold answers and correctness labels are not shown to target-model confidence instruments or to the semantic clustering model.

Judge sensitivity. Three swaps audit the semantic clustering judge: an auxiliary frontier-target SimpleQA panel, a full TriviaQA re-clustering with a stronger same-provider judge, and a cross-provider Qwen3-32B judge (Yang et al., 2025). Across the two full-panel swaps, 3.4–4.85% of partitions change; pooled C and fusion AUROC changes remain small with paired intervals crossing zero. Proper scores move slightly under the cross-provider swap, so ranking is stable while probability quality is not completely judge-invariant. All audits are gold-blind, and one documented false merge was resolved by a second judge rather than by correctness. Full per-row swaps, offsets, and anchor sensitivities are archived with the artifacts.

Truncation and parser audits. A post-hoc audit of legacy generation caps found a small number of confidence outputs recovered by a fallback parser from truncated text: 14 of 1,001 SimpleQA cells and 8 of 1,277 legacy multi-domain cells. Excluding every affected cell fail-closed changes each headline metric by at most .005 and no direction; all scaled-up collection records stop reasons with runaway-only caps. The scale-up parser accepts a bounded numeric Confidence line even when a model adds text after the requested final block; a stricter audit requiring the exact final block removes 257/69/76 GSM8K/TriviaQA/MMLU-Pro rows, and on those common subsets A+C remains best on all three metrics.

Encoder input encoding. Encoder inputs use pair encoding with a tail-preserving answer cap because 9.5% of raw MMLU-Pro rows exceed the 512-token limit; answer conclusions are verified preserved.

Label-audit refits. The conclusion-aware GSM8K audit changed 72 of 4,096 labels; every GSM8K compact method was refit and every exact encoder was retrained on the corrected labels rather than updated by scoring only.

Baseline-completion protocol. The comparators added in the final revision (the peers' verdict, two single verifiers, the released correctness model on the extension datasets and the frontier targets, the retrained encoder at the frontier, the strongest-single-peer bit, the verdict fusion ladder, and the faithful A+TU replication) were each preregistered before collection or analysis (analysis plans with SHA-256 prefixes `9f5186e5`, `ba73280b`, `6b922f1f`, and `51de7521`). Five dated amendments were recorded before results were read: an estimand-level parity gate for the correctness model in place of a row-level tolerance, an AUROC-gated parity cell for the encoder (later reproduced exactly after an input-order fix), a unicode-safe truncation fix in the correctness-model scorer, one budget-extended retry for judge calls that hit the token limit with paired exclusion of the residual rows (6 of 2,553 on GSM8K, 11 of 2,537 on MMLU-Pro, 3 of 2,551 on TruthfulQA, none on TriviaQA), and a parser accepting a markdown-emphasized probability line. Two further plans were frozen before their collection: the peers' verdicts on GPQA-Diamond and MedQA (prefix `0b8ad810`) and the frontier answers judged by the stale panel and by the refreshed panel (prefix `5b143af5`), the latter with one amendment (prefix `20f0aa2a`) recorded after collection and before analysis, adding a separate 16,000-token attempt and a fail-closed treatment as sensitivity

because the registered row exclusion drops harder rows (Appendix O). Registered gates were never widened; predictions that were not met are reported as such in Table 5. The judge collections comprised 64,210 canonical calls with no throttling failures, plus 14,176 for the two extension datasets and 16,832 for the frontier-target rows; the correctness model and the encoder ran on managed compute with a frozen parity cell each.

E HYPERPARAMETERS AND TRAINING SETTINGS

All settings below are read from the archived training code and job records; none were tuned on evaluation cells.

A+C fusion. Scikit-learn logistic regression, L2 penalty with $C = 1.0$ (the library default, never tuned), `lbfgs` solver, up to 3,000 iterations; features standardized to zero mean and unit variance. The scaler and classifier are refit inside every training split. There are no other free parameters: the feature set (six or three self-report instruments plus four agreement statistics) is fixed by the design matrix, so the estimator has at most eleven fitted parameters.

Exact peer baselines. Calib-text: BERT-base encoder with a linear head, AdamW, learning rate 10^{-5}, weight decay .01, batch size 16, up to 5 epochs with per-epoch validation selection, linear schedule without warmup, gradient clipping at norm 10. APRICOT (binary and cluster): DeBERTa-v3-base, AdamW with the learning rates and weight decays published by the original Bayesian search (binary: 1.406×10^{-5}, .0318; cluster: 3.369×10^{-5}, .0089), evaluated at the released step-budget fractions with validation selection. Maximum sequence length 512 everywhere; seed 123 for the primary run. The enlarged-training condition quadruples the peers' training rows and changes nothing else. The question-cluster intervals reported throughout cover question sampling. Encoder training is held fixed at one seed and is outside those intervals; the smallest fusion-versus-encoder AUROC margin is $+.191$.

Instruments and panel. The answer models and peers span five providers and include, among others, Qwen3 models (Yang et al., 2025) and DeepSeek-R1 (Guo et al., 2025), all accessed through standard inference APIs. The cheap-peer panels of Appendices G and L use Nova Lite, Qwen3-32B, and GPT-OSS-20B. Semantic clustering uses a gold-blind Haiku 4.5 judge; TriviaQA grading uses two independent gold-aware judges (Haiku 4.5 and Sonnet 4.5) with explicit adjudication of every disagreement. Behavioral resampling uses temperatures $\{.2, .7, 1.0\}$ with five samples each on the pilot and eight resamples per answer at scale; peer answers are collected once at the provider's default temperature and frozen. Semantic clustering and grading prompts, judge models, and adjudication rules are archived with the data; grading is deterministic on GSM8K and MMLU-Pro.

F TRANSFER EXPERIMENT DETAILS

Figure 2 (main text) summarizes the transfer comparison as effect sizes with uncertainty; the paragraphs below record the protocol and cost details behind it.

Labels and grading. GSM8K labels come from a conclusion-aware deterministic grader: an audit of the legacy any-number rule corrected 72 of 4,096 labels, one per question and collected model (512×8), of which 68 were decimal-format false negatives and 4 were incorrect intermediate-number matches. MMLU-Pro uses deterministic option grading; TriviaQA labels come from two independent gold-aware judges with explicit adjudication of every disagreement (13 in the first round, 52 at scale, 98.5% judge agreement). Because Sonnet 4.5 is both a grader and an evaluated target, the same four labelings (adjudicated primary, dispute-dropped, Haiku-only, Sonnet-only) were run on both judged panels. On TruthfulQA every conclusion is unchanged; the Haiku-only labeling, which removes the Sonnet judge from grading entirely, preserves the fusion's paired Brier improvement over self-report with a 95% interval excluding zero. On TriviaQA the Haiku-only labeling changes 53 of 2,537 panel-row labels and moves the fusion from .933/.038 to .899/.046 (AUROC/Brier) and the peer view from .922 to .885; the fusion stays above self-report (.769/.078) and above every retrained encoder on both metrics, and the Sonnet 4.5 target cell moves from .899 to .855. Direc-

tions do not depend on the shared judge; the TriviaQA levels do, by about .03 AUROC, and Table 1 reports the adjudicated labeling.

Exact baseline implementation. The APRICOT-cluster variant recomputes its cluster targets from source questions inside each split, preventing train–test leakage under our protocol. Because the published fixed-position Calib-n-joint head does not specify how to score an unseen model, we train with panel permutations and average all peer-omission and rotation views at test time (49 views on SimpleQA, 16 on the five-model panels). Each method selects its training duration on source-only validation questions and then refits from the pretrained checkpoint on all source questions. A native-regime sanity check comes from the stored training diagnostics: scoring the seen source models on held-out validation questions, Calib-n's intended use, the retrained head reaches median log loss .54/.29/.63 on GSM8K, TriviaQA, and MMLU-Pro, at or below each dataset's constant-prevalence reference, so the unseen-target adaptation does not cripple the method; the transfer penalty appears at the unseen-target step itself. In the enlarged-training condition, quadrupling the source data from the first 128-question round improved the encoders materially (GSM8K Calib-text .348 to .606, Calib-n .469 to .610) while every gap kept its direction and significance.

Peer separation under holdout. Agreement features are recomputed leave-one-model-out for every held-out target: training rows' C statistics come from clusterings that exclude the held-out model's answers, so the held-out model never enters any training row's peer panel, and its own evaluation rows take the remaining panel members as peers. Evaluation rows average the four agreement statistics over all four three-peer subsets of their panel, matching the three-peer geometry that training rows have once the held-out model is excluded.

Instrument-level cost detail. Each self-report instrument queries the expensive target, and the expected-value instrument emits a reasoning trace, which is where the bill sits: on the Sonnet-4.5 target one such call costs $.0069 against $.0006 for its four peer answers and $.0011 for the clustering call. Across the five panel targets the added bill spans $.004–$.009, always under a cent, and the panel is never its dominant term. The ratio to the answer's own generation cost is not a useful summary, because that denominator itself varies more than fiftyfold across the panel. Panel composition is therefore governed by separation (Section 4.3). Budget is not the limiting factor. Standalone, the instruments rank expected-value $>$ direct $>$ scored (.779/.745/.690 on GSM8K), so the elicitation ladder is itself informative and the scored instrument is redundant. On the same rows, the one-instrument fusion matches the full one: A2+C reaches AUROC .938/.933/.857 on GSM8K, TriviaQA, and MMLU-Pro against .938/.933/.859 for three instruments, and never gives up more than .002 on a proper score. A cascade that runs the cheapest instrument first and escalates only uncertain answers fails on the numbers: escalating half of GSM8K answers already costs 61% of the full per-answer bill and recovers only 20% of the AUROC gain (.745 to .784 against .938).

Risk–coverage detail. At a 5% observed error ceiling, the 92.5% TriviaQA coverage of the main text leaves 7.5% rather than 65% of answers for human review relative to the best retrained encoder. Pool error at 80% coverage on TriviaQA is .015 [.006, .026] against .054 for eight-sample self-consistency, .088 for Calib-n, and .110 for accepting everything. On the area under the risk–coverage curve, C dominates repeated self-sampling on all three 512-question panels (GSM8K .212 versus .318), but at particular coverages the sampling view can win (at 80% coverage on GSM8K its pool error is .375 against .438), and A+B+C attains the best GSM8K curve (.160).

G COMPARISON-LADDER CONTROLS

This appendix is a baseline roster. Each rung instantiates a published black-box signal family under the identical frozen-answer protocol. Verbalized elicitation instruments (expected-value, direct, scored) represent the self-evaluation family; the eight-resample consistency signals represent self-consistency and semantic entropy; the paraphrase pilot probe below represents the perturbation family; the C statistics represent cross-model agreement; text-only supervision and source prevalence are shortcut controls; and Section 3's exactly retrained encoders represent the supervised calibrators. Across the four core panels, no standalone family reaches the fusion.

Table 9 reports the lower rungs of the comparison ladder on the SimpleQA panel. The text-only controls run under the same simultaneous held-out model and question block as the headline table; so does the nested quadratic comparison, which selects its regularization inside grouped inner cross-validation. The classifier grid uses the leave-one-model-out split with configurations fixed before evaluation. The results have two readings: text alone carries no usable signal under this estimand, and added classifier capacity on the identical A+C features buys nothing (the tree ensembles trade a small AUROC movement for materially worse proper scores).

Split	Model	AUROC	Brier	Log loss
double holdout	TF–IDF question	.380	.185	.564
	TF–IDF answer	.445	.185	.566
	TF–IDF question+answer	.447	.181	.553
	answer style features	.516	.177	.546
	A+C (reference)	.824	.136	.423
double holdout	quadratic minus linear A+C	$+.0002$	$-.0013$	$-.0076$
	95% CI	$[-.013, .012]$	$[-.007, .005]$	$[-.025, .009]$
leave-one-model-out	logistic (reference)	.829	.132	.414
	quadratic logistic C=0.1	.835	.128	.401
	boosting depth 2 / 3	.828/.826	.130/.132	.408/.414
	random forest	.837	.158	.485
	radial-basis support-vector classifier	.803	.140	.439

Table 9: **Text supervision carries no usable signal under double holdout; added predictor capacity brings no consistent gain on A+C.** The controls use the SimpleQA panel, with classifier variants evaluated on identical A+C features; the A+C reference is recomputed inside the ladder pipeline, whose deterministic peer selection gives .824 against the headline pipeline's .823.

Same-feature capacity on the 512-question panels. The SimpleQA control above holds on its panel. A preregistered extension runs the same head grid on identical A+C features under the published double-holdout folds of the three 512-question panels. Its preregistered predictions failed and are kept as registered: on GSM8K a depth-2 boosted head improves all three metrics over the linear reference (AUROC .938 to .962, Brier .113 to .082, log loss .332 to .271), while on TriviaQA and MMLU-Pro the linear head remains best on every metric. The quadratic head, the most controlled capacity increment, never wins. Added final-head capacity on the same features is therefore dataset-dependent, which is how the main text scopes the measurement-versus-predictor contrast. An exploratory sweep (not preregistered) over the head's regularization strength and penalty, class weights, four feature maps, and a small gradient-boosted tree, with every choice made by source-only cross-validation inside each training cell, moves pooled A+C by less than .025 on GSM8K and MMLU-Pro, and no configuration, even one selected on the test rows, reaches the peers' verdict on either dataset.

Encoders given the measurements. A preregistered paired control fits a converged L2-logistic head, on the same double-holdout folds and rows, to the frozen Calib-text encoder's CLS representation, to the seven A+C measurements, and to both together; the measurements-only head reproduces the linear fusion within .004 AUROC on every dataset. Table 10 shows that the frozen text representation keeps only within-environment ordering, at the level of the fine-tuned encoder (Appendix H), and that a head given the measurements alongside it recovers the fusion on GSM8K and TriviaQA. The registered prediction of at least 90% of the gap on every dataset failed on MMLU-Pro (87%), where the 768 text dimensions cost .02 AUROC under the source-selected penalty; per cell, adding them never helps in more than 6 of 25 cells. Warm-starting the head from the fusion reaches the same optimum.

Sampling and perturbation controls. Table 1 reports the eight-resample B view. A preregistered paraphrase baseline reaches the same conclusion on the three 512-question panels: paraphrase consistency remains below C, and fusing the implementable sampling and paraphrase branches draws level with C only on GSM8K. Meaning-preservation is itself a limitation: an adversarial three-validator gate certifies only 27.4% of audited nondegenerate paraphrases, while two blinded human

Dataset	Fine-tuned Calib-text	Frozen Calib-text	Frozen Calib-text + A/C	Linear A+C	Gap closed
GSM8K	.606	.601	.951	.938	104%
TriviaQA	.620	.662	.932	.933	100%
MMLU-Pro	.669	.665	.835	.859	87%

Table 10: **A fitted head given the measurements nearly recovers the fusion; the text representation adds no reliable gain.** Entries are pooled AUROC under the paired double-holdout protocol; "gap closed" is the Calib-text-to-linear-A+C gap closed by the frozen-Calib-text + A/C head.

annotators show that this gate is a conservative lower bound and that its hardest rejections are genuinely ambiguous. Using all rows is therefore the paraphrase-favorable comparison.

The five-family perturbation composite. We then ran the implementable black-box families of Pedapati et al. (2025) (stochastic decoding, paraphrase, stopword removal, entity-frequency amplification, and sentence permutation, which needs a multi-sentence stem and enters the composite only on GSM8K; their sixth family, split-response consistency, needs a multi-part frozen response and an entailment model and was not run) with matched regenerations on the rows where every used family is valid (2,042 GSM8K, 2,193 TriviaQA, 1,791 MMLU-Pro rows; A+C on those rows reproduces its full-row AUROC to within .006). Four predictions were preregistered. The one that held everywhere is the one the paper needs: the perturbation composite stays below A+C on every metric and dataset (AUROC .901 against .937 on GSM8K, .848 against .928 on TriviaQA, .750 against .857 on MMLU-Pro; paired intervals exclude zero). Three failed, all on GSM8K, and they locate the exception: there the perturbation fusion beats the peer view alone (.901 against .826), pairing it with self-report edges the A+C fusion (.952 against .937), and sentence permutation alone beats stochastic decoding. On TriviaQA and MMLU-Pro every family stays below the peer view. An exploratory, unregistered addition of the perturbation composite to A+C gives .971/.937/.861, so the two information sources are complementary where the target's behavior is informative.

A released correctness model. The Generalized Correctness Model of Xiao et al. (2025) is an 8B language model fine-tuned on the correctness histories of eight source models to predict, from the question and answer text, whether an unseen model's answer is correct. It is the strongest released text comparator we know of and it has no access to our labels, so we applied the released adapters zero-shot to our Table-1 rows (the TriviaQA adapter on GSM8K and TriviaQA, the MMLU adapter on MMLU-Pro; five predictions preregistered). It transfers far better than the retrained encoders (AUROC .808/.863/.714 against the best encoder's .630/.671/.669) and, unlike them, keeps positive cross-model level alignment, yet it stays below the peer view alone at every point estimate and below A+C by .07–.15 with intervals excluding zero. Its MMLU adapter saw 212 of the 512 MMLU-Pro stems in training; on the clean 300-stem subset it scores .752 against A+C's .877. Its 4,096-dimensional final representation fitted with our protocol's head ranks below the seven measurements everywhere and, appended to them, lowers AUROC by .013–.066; the frozen-encoder control above likewise finds no reliable gain from text. One prediction failed: on GSM8K its per-cell AUROC is below its pooled AUROC, the signature of level alignment rather than the encoders' failure. Adaptation and inference costs are separate questions; this paper claims no inference-cost advantage. Full contrasts, human paraphrase labels, and disagreement buckets are archived.

Peer verdicts and single verifiers. Section 3 reads the same peers two ways. J follows C's leave-one-model-out geometry: the held-out model's verdicts are removed from every training row, its own rows are judged by the four others (mean over the four three-judge subsets on the letter-answer datasets), and count-based gates assert both. Table 11 gives the full fusion ladder for the four peers and for two single verifiers, all views passed through the same source-trained head. Two registered predictions about this comparison were not met: that the peers' verdict would rank below their agreement count, and that A+J would rank below A+C; on GSM8K and MMLU-Pro both orderings reverse, and on TriviaQA the two reads are level. A separately registered prediction, that adding the count to a verdict never hurts, held: A+C+J ranks at or above A+J in every cell, and at or above the better of A+C and A+J in 15 of 16 (the exception is Qwen3-32B on MedQA, by less than .001); the registered pattern for the size of that increment across datasets was not met. A later plan extended the peers' verdict to GPQA-Diamond and MedQA, whose letter labels make all four non-target peers label-independent judges, with the registered predictions that the verdict beats

the count and A+J beats A+C on both; both held (GPQA .845 against .704, +.141 [+.091, +.197]; MedQA .929 against .866, +.062 [+.033, +.094]; A+J against A+C +.083 [+.049, +.118] and +.037 [+.017, +.058]), on the 947 and 2,551 rows where every peer returned a verdict. The registered prior that the verdict advantage would be smaller on GPQA, where the panel is right only about half the time, was wrong in direction: the gain there matches GSM8K's. On TruthfulQA, where the agreement count is weakest, the peers' verdict reaches .785 against .646 for the count. A single verifier is a different object from the panel: Qwen3-32B's stated probability ranks at .979/.816/.766 (GSM8K/TriviaQA/MMLU-Pro) before any head, below the agreement count on TriviaQA and MMLU-Pro, and its raw probabilities are poorly calibrated: Brier .030/.115/.195 raw, .032/.085/.178 after a two-parameter Platt map fitted under the same double holdout, against .147/.040/.143 for the count. Haiku 4.5 as a single verifier reaches Brier .011 and .117 after the same map on GSM8K and MMLU-Pro. A third registered prediction, that agreement with the strongest single peer would match the four-peer count, was not met either: the one-peer bit ranks at .796/.848/.861 against the count's .837/.922/.836, below it on GSM8K and TriviaQA and above it on MMLU-Pro, with Sonnet 4.5 the strongest peer in 20 of 25 training cells on each dataset. Costs are reported without netting them against accuracy. The agreement count needs the peers' answers, which the protocol already collects; the peers' verdict needs one further call per peer (4 calls and about 1.3k tokens per GSM8K row, 4 calls and 2.3k tokens per MMLU-Pro row); a single verifier needs one call (Qwen3-32B about 0.3k and 0.5k tokens per row, Haiku 4.5 about 0.4k and 0.8k with its reasoning output). Choosing among these reads is a labeled decision (Section 6); the paper reports the ladder and does not automate the choice.

Judge source	Dataset	A	C	J	A+C	A+J	A+C+J	excl.
Four peers	GSM8K	.804	.837	.993	.938	.994	.996	6
Four peers	TriviaQA	.774	.922	.908	.933	.910	.938	0
Four peers	MMLU-Pro	.757	.835	.881	.859	.882	.884	11
Four peers	TruthfulQA	.749	.646	.785	.764	.799	.799	3
Four peers	GPQA	.707	.704	.845	.763	.846	.851	37
Four peers	MedQA	.764	.866	.929	.892	.928	.929	9
Qwen3-32B	GSM8K	.804	.837	.972	.938	.983	.990	0
Qwen3-32B	TriviaQA	.774	.922	.747	.933	.846	.934	0
Qwen3-32B	MMLU-Pro	.756	.836	.718	.859	.809	.866	0
Qwen3-32B	GPQA	.719	.704	.559	.770	.728	.771	0
Qwen3-32B	MedQA	.765	.867	.725	.892	.807	.891	0
Qwen3-32B	TruthfulQA	.749	.646	.702	.764	.775	.779	0
Haiku 4.5	GSM8K	.804	.837	.992	.938	.994	.996	0
Haiku 4.5	MMLU-Pro	.756	.836	.887	.859	.887	.892	0
Haiku 4.5	GPQA	.719	.704	.841	.770	.841	.847	0
Haiku 4.5	MedQA	.765	.867	.905	.892	.910	.924	0

Table 11: **Adding a verdict view to the same source-trained head.** Pooled AUROC under the transfer protocol of Section 3; J is the judge's stated probability that the frozen answer is correct, from the four peers (mean over the peers that returned a valid probability) or a single verifier, passed through the same head as every other view. The last column counts rows excluded because a peer judge returned no usable probability after one extended retry (Appendix D); on those datasets every column is refit on the remaining rows. Haiku 4.5 also served as a labeler elsewhere in the paper and is reported as a sensitivity arm.

Panel size and diversity. Peer information improves monotonically with panel size on SimpleQA, the three 512-question panels, and an auxiliary frontier-target SimpleQA panel. On SimpleQA, adding five diverse peers moves AUROC from .663 to .817 and Brier from .166 to .135; the frontier panel shows a similar ranking gain, while high-coverage selective-risk intervals remain inconclusive. Provider diversity matters more as the panel grows. The complete subset curves and coverage-dependent uncertainty are archived; the main configuration uses four peers.

Cheap peers on the standard benchmarks. Replacing the frozen panel with three roughly twenty-times-cheaper peers (Nova Lite, Qwen3-32B, GPT-OSS-20B) gives a dataset-dependent trade: the cheap-panel fusion improves on GSM8K (A+C AUROC .954 against .938) and loses on TriviaQA (.846 against .933) and MMLU-Pro (.812 against .859). The hard-regime companion, where cheap peers degrade the agreement view, is in Appendix L.

	SimpleQA (prev. .227)			GSM8K (.514)			TriviaQA (.890)			MMLU-Pro (.660)		
Method	AUC	Br	LL	AUC	Br	LL	AUC	Br	LL	AUC	Br	LL
A+C (minimal)	**.823**	**.136**	**.423**	.938	.113	.332	.933	.038	.142	.859	.128	.414
A+B+C (+8 target resamples)	N/A	N/A	N/A	**.968**	**.070**	**.227**	**.938**	**.037**	**.139**	**.862**	**.125**	**.407**
C only	.797	.138	.434	.837	.147	.462	.922	.040	.151	.836	.143	.450
A only	.661	.166	.512	.804	.127	.465	.774	.067	.251	.756	.175	.535
B only (8 target resamples)	N/A	N/A	N/A	.805	.163	.509	.726	.083	.290	.650	.189	.567
Source prevalence	.333	.179	.547	.299	.275	.744	.378	.099	.351	.345	.234	.661
Calib-text	.473	.182	.557	.606	.262	.732	.620	.097	.356	.669	.212	.617
APRICOT-binary	.581	.261	.716	.630	.258	.711	.511	.239	.670	.555	.249	.691
APRICOT-cluster	.387	.181	.552	.370	.281	.762	.478	.099	.355	.467	.232	.656
Calib-n-joint	.393	.180	.551	.610	.250	.704	.671	.100	.381	.618	.220	.636

Table 12: Complete transfer comparison (full three-metric version of Table 1). AUC is AUROC; Br is Brier score; LL is log loss. Every reported cell uses the same double-holdout protocol; view-specific missing calls produce the exact denominators reported in the archived predictions. B and A+B+C are N/A on SimpleQA because its available B result is a one-target pilot; the exact encoder rows provide the complete supervised comparison. Proper scores are primary; a random ranking has expected AUROC .500. Pooled AUROC below .500 is driven by the between-cell level misalignment analyzed in the text.

H POOLED VERSUS PER-CELL DECOMPOSITION

Across the sixteen A+C-versus-encoder comparisons of Section 3, every 95% question-cluster bootstrap interval excludes zero; the smallest AUROC margin is $+.191$ $[+.155, +.226]$, on MMLU-Pro.

The pooled margins hide environment-level exceptions. In 6 of the fifteen 512-question environments one component score ranks above the fusion by more than .005 AUROC. On the strongest target of every panel, Sonnet 4.5, the single direct-confidence report reaches .953/.932/.766 against the fusion's .930/.899/.718 on GSM8K/TriviaQA/MMLU-Pro, and the three-instrument self-report view does so on GSM8K and MMLU-Pro: this is the separation mechanism of Section 4.3 already inside the source generation, where a target far above its panel gains nothing from agreement. At the other end, on Nova Pro, a bottom-two target on every panel, the peer view alone beats the fusion on GSM8K (.975 against .846) and MMLU-Pro (.908 against .887); there the self-report link fails (Appendix I) and the fusion inherits its noise.

Table 13 decomposes every pooled AUROC of the 512-question panels. The pooled reading tests one shared deployment scale across models; the per-cell reading tests ranking within one target environment. We report mean per-cell AUROC over the twenty-five heldout-model by question-block cells and cross-cell level alignment (Pearson correlation between a method's mean score in a cell and the target's actual accuracy there). Individual encoder cells range below .30 (Calib-n on GSM8K reaches .179). Every entry whose pooled AUROC falls below the .500 coin flip (APRICOT-cluster on all four datasets, Calib-text and Calib-n-joint on SimpleQA, and the source-prevalence constant) has a positive mean within-cell ordering; cross-model level misalignment is what pushes the pooled number below one half. The constant source-prevalence baseline makes the mechanism exact: within a cell it is a coin flip (.500) with level anti-correlation -1.00 by construction, pooling to .299–.378, while a genuinely random ranking pools to .500 because its score distribution is identical in every cell and pays no level penalty. A+C's own per-cell AUROC never falls below .659 in any of the seventy-five 512-question cells, and the C view pairs strong per-cell ordering (.920 on GSM8K) with negative level alignment ($-.76$). The mechanism behind the strongest anti-correlations is visible in the architectures: APRICOT-cluster regresses source-panel cluster accuracy, so a target that beats the source panel receives systematically lower scores and a weaker one higher scores (Pearson $-.99$ on GSM8K; $-.86$ to $-.95$ for the three SimpleQA encoder entries below .5, whose per-cell means are .51–.56). The scalar encoders see only question and answer text; Calib-n-joint also sees peer-answer text but no target-model labels or self-report.

Encoder slope anatomy. The encoders' pooled fitted slopes of .34–.69 understate the within-cell relationship because they also absorb the cross-model level misalignment quantified above; repairing that misalignment cell by cell would require labeled anchors from every deployed model–dataset pair, which is exactly what the transfer setting denies. On the pilot, post-hoc Platt scaling

on the evaluation slice itself returned the encoders to the constant-prediction baseline at that slice's prevalence.

	GSM8K			TriviaQA			MMLU-Pro		
Method	pooled	cell	level	pooled	cell	level	pooled	cell	level
A+C	.938	.956	+.48	.933	.959	+.91	.859	.852	+.86
C	.837	.920	−.76	.922	.940	+.85	.836	.826	+.40
A	.804	.917	+.56	.774	.864	+.35	.756	.781	+.74
Source prevalence	.299	.500	−1.00	.378	.500	−1.00	.345	.500	−1.00
Calib-text	.606	.629	−.33	.620	.650	−.87	.669	.618	+.83
APRICOT-binary	.630	.639	+.29	.511	.644	−.47	.555	.591	−.13
APRICOT-cluster	.370	.601	−.99	.478	.591	−.23	.467	.576	−.98
Calib-n-joint	.610	.690	−1.00	.671	.693	−.94	.618	.615	−.08

Table 13: **Text encoders retain within-cell ordering and lose cross-cell level alignment.** The table reports pooled AUROC, mean per-cell AUROC ("cell"), and cross-cell level alignment ("level"; Pearson correlation of per-cell mean score with per-cell accuracy) on the 512-question panels. The parenthesized value under each dataset name is the correct-answer prevalence ("prev.") of its evaluation rows.

I DIRECT CALIBRATION DISPLAY

Proper scores mix discrimination with calibration, so Table 14 shows the probability scale of the zero-shot A+C fusion directly on the 512-question panels: evaluation accuracy, mean predicted probability, their difference (pooled calibration-in-the-large, CITL), class-conditional mean confidences, and the cross-fitted recalibration headroom, defined as the change in log loss from an out-of-fold affine (Platt) recalibration by question block. Positive headroom means a two-parameter repair would still help; negative means it hurts. Pooled CITL is close to zero on TriviaQA and MMLU-Pro, where recalibration indeed hurts, and $+.018$ on GSM8K with a small positive headroom, consistent with fitted links close to the affine refinement of Section 4.1 (pooled slopes $.67/1.00/1.06$). Question-cluster bootstrap 95% intervals for the pooled CITL are $[+.009, +.027]$, $[-.009, +.011]$, and $[-.019, +.021]$. The class-conditional confidence separations (mean confidence on correct minus on wrong answers: $.648/.609/.413$) measure discrimination and should not be interpreted as calibration.

Pooled level alignment is a weak certificate on its own, for two reasons. First, it is nearly free: any method whose cross-fitted intercept tracks the source base rate attains it, and indeed the source-prevalence constant reaches pooled CITL below .001 on all three panels. Second, pooling can hide environment errors that cancel. Table 15 therefore reports the fifteen environments separately. The pooled GSM8K value of $+.018$ decomposes into $+.333$ on the Nova Pro–GSM8K link-failure cell and $-.263$ on the strongest model, which the agreement view underestimates because peers cannot see that the target outruns its panel; the other three GSM8K cells sit within .031. The discriminating comparison is the median absolute per-environment CITL: .030 for A+C against .087 for C alone and .122 for source prevalence. Pooled CITL primarily describes aggregate level alignment; A+C's answer-specific value is established by its per-environment level accuracy together with the discrimination and proper-score results of the main text.

Dataset	Acc.	Mean p	Pooled CITL	$\mathbb{E}[p \mid Y{=}1]$	$\mathbb{E}[p \mid Y{=}0]$	ΔNLL recal.
GSM8K	.514	.532	+.0178	.847	.199	+.025
TriviaQA	.890	.891	+.0002	.957	.348	−.007
MMLU-Pro	.660	.660	−.0001	.800	.387	−.001

Table 14: Zero-shot A+C probability scale, pooling out-of-fold predictions across all held-out target-model and question-block cells within each dataset. Pooled CITL can conceal environment-specific errors that cancel; per-environment values are in Table 15. ΔNLL recal. is the log-loss change from out-of-fold affine recalibration (positive: recalibration would help; negative: it hurts).

Recalibration deltas and comparators. Out-of-fold affine recalibration by question block changes pooled log loss by only $+.025/-.007/-.001$ on GSM8K, TriviaQA, and MMLU-Pro,

Dataset	Target	Acc.	Mean p	CITL	Slope	Intercept	NLL
GSM8K	Maverick	.417	.407	$-.010$	1.37	$+.51$	.049
	Mistral	.361	.359	$-.002$	1.38	$+.13$	.043
	Nova Pro	.375	.708	$+.333$	0.54	-1.68	1.023
	Qwen3	.487	.518	$+.030$	1.72	-1.18	.119
	Sonnet 4.5	.930	.667	$-.263$	1.36	$+3.62$	.424
TriviaQA	Maverick	.880	.874	$-.006$	1.11	$+.05$	.132
	Mistral	.908	.904	$-.005$	0.99	$+.13$	.136
	Nova Pro	.850	.874	$+.024$	1.01	$-.58$	.173
	Qwen3	.870	.888	$+.018$	1.00	$-.54$	.146
	Sonnet 4.5	.943	.913	$-.030$	0.96	$+.91$	.123
MMLU-Pro	Maverick	.768	.688	$-.080$	0.98	$+.65$	.398
	Mistral	.547	.600	$+.054$	1.05	$-.50$	.397
	Nova Pro	.527	.619	$+.091$	1.24	$-.82$	.451
	Qwen3	.612	.685	$+.073$	1.41	-1.11	.377
	Sonnet 4.5	.847	.708	$-.139$	0.76	$+1.15$	.447

Table 15: Per-environment probability scale of zero-shot A+C: calibration-in-the-large, calibration slope and intercept (logistic fit of correctness on the logit score), and log loss (the NLL column) for each held-out target model. Median/mean/maximum absolute CITL across the fifteen environments: $.030/.077/.333$ for A+C, against $.087/.105/.473$ for C only and $.122/.130/.520$ for source prevalence.

negative where it hurts; pooled alignment is nearly free in the sense that the source-prevalence constant also attains it, which is why the per-environment display above is the informative summary. The median absolute per-environment level error is $.030$ for A+C, $.087$ for the C view alone, and $.122$ for source prevalence.

The one failing link. The A+C link for Nova Pro on GSM8K has a large intercept correction and a slope different from one in both the 128-question round and the 512-question pooled out-of-fold fit. It resembles an intercept-dominated label shift, but it is not pure label shift because the slope also changes; pooled zero-shot CITL across the three panels is $+.018/.000/.000$.

Within-fold curvature diagnostic. For each environment we compare the affine link with a nested quadratic-in-log-odds alternative on each fold's out-of-fold scores and sum the five one-parameter likelihood-ratio statistics. The asymptotic χ^2_5 reference flags three of the fifteen 512-question environments at the $.01$ level; a 10,000-replicate conditional parametric bootstrap at the fixed out-of-fold scores flags two of the three, with the third at an estimated $p = .014$. Both references are diagnostics rather than exact tests because the fold-specific score models share training rows. The largest departure, Nova Pro on GSM8K, has pooled LR statistic 81.7 and is flagged by either reference; on TriviaQA no environment rejects, consistent with fitted links near the identity there. In the flagged Nova Pro–GSM8K cell, flexible isotonic recalibration reaches cross-fitted log loss $.377$ against $.492$ for the affine repair.

Direct reversal test. A reversal test on the same out-of-fold scores (ten equal-count bins per environment, question-cluster bootstrap) detects a local decrease of the correctness rate in two of the fifteen environments: the Nova Pro GSM8K cell above and a three-answer dip at the saturated top of the Sonnet 4.5 GSM8K scale (bin rates 1.00 to $.94$). The other thirteen show none at this resolution, which bounds but does not certify the monotone link. Held-out isotonic recalibration is worse than the affine link in fourteen of fifteen environments (median $+.020$ log loss). The decision cost of the two detected reversals is measurable: accepting the top half of answers by the score instead of by an out-of-fold binned estimate of $\mathbb{P}(Y{=}1 \mid S^*)$ costs $.051$ $[.012, .094]$ of accepted-pool error in the Nova Pro GSM8K cell and $.042$ $[.022, .068]$ at 80% coverage, $.012$ $[.000, .027]$ at half coverage in the Sonnet 4.5 GSM8K cell and nothing at the coverages a strong model would run at, and within estimation noise of zero in the other thirteen environments (question bootstrap; the binned estimate is itself noisier than the score at high coverage).

J SOURCE-SELECTED THRESHOLD STRESS TEST

Pooled proper scores do not by themselves establish that a decision threshold transfers. We therefore run a stricter source-only diagnostic on the corrected 512-question panels. For each outer (target model, question fold) cell, the final estimator fits only the other models and other question folds. Its threshold is chosen from held-out predictions of an inner question-fold cross-validation run entirely within those outer-source rows; neither target-model labels nor the outer question block enter feature standardization, estimator fitting, or threshold selection. Tied scores enter as a complete level. The chosen threshold is the most permissive one whose empirical accepted-pool error on the inner source predictions is at most the requested level. This diagnostic measures transfer and provides no finite-sample risk guarantee. Per-environment Wilson intervals and all five fold-level thresholds are archived with the results.

Table 16 shows a mixed result. At a 5% requested risk, the C-only point estimates meet the target in all five GSM8K environments and four of five TriviaQA environments, whereas A+C meets it in four and one: the target self-report improves average proper scores but can hurt a strict shared threshold when its model-relative level shifts. In particular, the Nova Pro–GSM8K link failure raises the pooled A+C accepted risk to .201, while C alone stays at .025. Neither method supports a useful 5% threshold on MMLU-Pro; at 10%, A+C is close in aggregate (.103) but only two of five model environments meet the requested level. Source prevalence supplies no usable strict threshold. Thus a common scale must be tested empirically. Pooled calibration does not imply one, and target-local anchors are a plausible repair rather than a prerequisite for the zero-label estimator.

Method	Dataset	Requested	Env. met	Pooled risk	Coverage	Worst risk
A+C	GSM8K	.05	4/5	.201	.502	.585
C	GSM8K	.05	5/5	.025	.284	.041
A+C	TriviaQA	.05	1/5	.055	.931	.082
C	TriviaQA	.05	4/5	.038	.903	.055
A+C	MMLU-Pro	.10	2/5	.103	.404	.113
C	MMLU-Pro	.10	2/5	.105	.251	.182

Table 16: **A source-selected risk threshold does not transfer reliably across environments.** "Env. met" counts target-model environments whose point-estimate accepted risk does not exceed the requested level; zero-coverage environments do not count as meeting it.

K CROSS-DATASET AND CROSS-MODEL STRESS TEST

The primary protocol transfers across models and questions within a dataset. To test a stronger boundary, for each target dataset D_t and model M_t we fit the common seven-feature estimator only on rows with $D \neq D_t$ and $M \neq M_t$, freezing the source-only scaler and logistic coefficients. The target dataset contributes no prevalence, labels, hyperparameters, or feature normalization; the three self-report and four agreement features have the same definition in every dataset. This gives fifteen held-out dataset–model environments.

Table 17 shows that the ordinal component transfers cleanly: A+C exceeds chance in all fifteen environments, all fifteen question-bootstrap AUROC lower bounds exceed .5, and the minimum point estimate is .721. Its mean AUROC is nearly unchanged from the within-dataset per-model reference (GSM8K $+.018$, TriviaQA $+.026$, MMLU-Pro $-.000$). The probability level is less stable: compared on the same target rows, mean A+C log loss changes by $-.092$, $+.035$, and $+.099$, respectively. Thus cross-dataset evidence supports the paper's separation rather than a universal calibration claim: answer ordering can survive a task change while the cardinal link still needs target-local repair.

Two qualifications frame the comparison. First, it is not source-budget matched: each cross-dataset fit sees about 4,060 source rows against roughly 1,630 per within-dataset fold, so the near-zero ΔAUROC establishes that the ranking survives the dataset change. It does not establish that cross-dataset training is intrinsically superior. Second, the environment-level probability errors are not a base-rate story: the constant source-prevalence predictor has mean environment |CITL| of $.335/.304/.144$, which the fused score reduces to $.094/.040/.109$, and the largest residual level er-

Target dataset	A AUROC	C AUROC	A+C AUROC	A+C Brier	A+C NLL	ΔAUROC
GSM8K	.917	.927	.972	.071	.240	+.018
TriviaQA	.875	.936	.955	.046	.177	+.026
MMLU-Pro	.782	.829	.846	.146	.513	−.000

Table 17: **Ordering survives double shift while probability scale remains environment-dependent.** Values average over five held-out target models; ΔAUROC compares A+C with its within-dataset per-model mean on the same target rows. This is a boundary stress test. The primary estimand remains within-domain transfer.

ror sits on GSM8K/Sonnet 4.5 ($-.296$) even though that cell has the smallest prevalence gap in its dataset. The unrepaired component is target-model-relative, the same term Section 4.3 shows source data cannot supply.

Few-anchor repair. We next ask how many labeled target examples repair the remaining probability scale. For each environment we draw $n \in \{5, 10, 20, 40\}$ anchors (500 paired draws per size, shared across configurations), fit either a smoothed intercept-only shift or a positive-slope affine logit link on the anchors alone, and evaluate on the held-out remainder. Table 18 reports the mean improvement in NLL.

Dataset	Anchors	Intercept ΔNLL	Affine ΔNLL	Single-class rate
GSM8K	5 / 10 / 20 / 40	.050 / .063 / .070 / .074	.029 / .065 / .090 / .104	.209 / .112 / .046 / .009
TriviaQA	5 / 10 / 20 / 40	−.036 / −.010 / .001 / .009	−.031 / −.012 / −.010 / −.002	.563 / .330 / .134 / .020
MMLU-Pro	5 / 10 / 20 / 40	−.000 / .038 / .063 / .078	.004 / .034 / .065 / .092	.179 / .055 / .008 / .000

Table 18: Cross-dataset few-anchor repair of A+C, averaged over five target models and 500 paired anchor draws. Positive ΔNLL favors repair. Intercepts use Beta$(1/2, 1/2)$ (Jeffreys) half-count prevalence smoothing; affine links have positive slope and the estimator's fixed regularization. Single-class draws fall back to the smoothed intercept.

The repair pattern is heterogeneous rather than a uniform dataset offset. On GSM8K the gain concentrates where the level error lives: the Sonnet 4.5 cell gains .276 NLL from five intercept anchors, while already-aligned cells are slightly harmed. TriviaQA links are near identity and small anchor sets usually hurt. At forty anchors an isotonic repair is worse on average on all three datasets ($-.051/-.246/-.305$ NLL for A+C). When repair is warranted, the affine link is therefore the practical flexible option; the intercept-only fit can have lower variance with very small anchor sets. Finally, the deployable one-instrument configuration transfers too: A2+C stays within .01 AUROC of the seven-feature fusion in 13 of 15 held-out environments with no tuning of any kind.

The cost of choosing a repair. The table above prices each repair given the choice; a deployed operator must also make the choice from the same labels. A prespecified policy therefore selects among nine configurations (raw self-report, or the paper's frozen out-of-fold A or A+C scores, each unrepaired, intercept-repaired, or affine-repaired) by five-fold cross-validation inside the audit set alone, refits the selection, and is evaluated on the disjoint remainder (500 draws per environment) at audit sizes of 10, 20, 40, 80, and 120 labels. The result has two regimes. Where the transferred scale is badly off, the policy helps at every budget with harm probability near zero. On GSM8K Sonnet 4.5 it moves median NLL from .424 to .124 at twenty labels and .105 at 120, and on GSM8K Nova Pro from 1.023 to .563 and .493. Where the unrepaired fusion is already near the oracle, small audits are fooled because a small, mostly-correct sample favors the sharpest score. The probability of being worse than the unrepaired default by more than .05 NLL reaches .812 at ten labels and .622 at twenty, exceeds .25 in ten of fifteen environments at ten labels, and falls to at most .106 at 120 labels, where fourteen of fifteen environments sit at or below .05. At 80 and 120 labels the policy's median NLL is within .019 and .013 of the best fixed rule (always raw, always A, always A+C, always intercept-repaired, or always affine-repaired) in every environment. Three registered predictions failed and are kept as registered: harm at 120 labels was not at or below .10 everywhere (GSM8K Mistral .106), the MMLU-Pro Sonnet 4.5 cell did not gain .05 NLL at twenty labels, and the within-.01 agreement with the best fixed rule held in eleven environments against the twelve required. An earlier avoid-harm prediction at twenty labels also failed. In two TriviaQA cells the policy's median NLL exceeds the always-affine comparator's, Maverick by .0006 and Sonnet 4.5 by

.0117, the latter a genuine repair the policy fails to capture. The measured advice matches the main text: audit the environment-local scale first, repair only where that audit finds a shift, and treat the choice of repair as a decision that needs an audit of roughly a hundred labels unless the shift is large.

Every score at the same audit budget. The policy above chooses among scores; a reviewer may instead ask how each score fares when all receive the same target labels. For each environment and draw (500 per environment), one permutation supplies nested audit sets of 20, 80, and 120 labels; every score (raw self-report, A, C, A+C, and the released correctness model) selects its own repair by the same five-fold rule on those labels, refits, and is evaluated on the same held-out remainder. A+C has the lowest median NLL in 10 of 15 environments at twenty labels and 9 of 15 at 80 and 120; where it loses, it loses to one of its own components (C on the weak-target cells, A on the two strongest targets), never to raw self-report or to the correctness model. The correctness model's median NLL stays above A+C in all 15 environments at every budget and below A in only 4 at 120 labels; on GSM8K it sits at .215–.604 against .044–.493 for A+C, since repair cannot rescue a weak ranking. On the MMLU-Pro subset free of its training overlap the ordering is unchanged, with one near-tie on Sonnet 4.5 (.371 against .368). For A+C, moving from twenty to 120 labels lowers median NLL by at most .025 in 13 of 15 environments; the two exceptions are the known scale-shift cells (GSM8K Nova Pro .065, MMLU-Pro Sonnet 4.5 .054).

Zero-shot MedQA transfer. The MedQA row of Table 3 holds out target models within MedQA under the protocol of Section 3. Pushing the same construction to a fully unseen domain changes almost nothing: for each held-out MedQA target model, the estimator is fit only on the GSM8K, TriviaQA, and MMLU-Pro rows of the four other models and scored on the target with no refit and zero MedQA labels, and this zero-shot fit nearly matches an in-domain MedQA fit while the probability level still moves (pooled AUROC .899 against .892, pooled level error $+.016$). Table 19 gives the per-model values.

Target	Acc.	AUROC	CITL	Brier
Sonnet 4.5	.904	.901	$-.067$	.074
Qwen3	.834	.912	$+.047$	.087
Maverick	.846	.909	$-.019$	.071
Mistral	.738	.944	$+.087$	.095
Nova Pro	.785	.892	$+.032$	.114
Pooled	.821	.899	$+.016$	.088

Table 19: Zero-shot cross-domain transfer to MedQA: A+C fit only on GSM8K, TriviaQA, and MMLU-Pro rows from the four non-held-out models, evaluated on each held-out MedQA target. CITL is calibration-in-the-large (mean predicted probability minus accuracy), in probability units.

L THE GPQA-DIAMOND, MEDQA, AND TRUTHFULQA ROUNDS

Three later panels extend the study: GPQA-Diamond (198 questions, five models, 984 complete rows), MedQA (512 questions, five models, 2,560 rows), and TruthfulQA (512 questions, five models, 2,551 rows), collected under the unchanged fixed-answer protocol, dual gold-aware judging with explicit adjudication on TruthfulQA, and the same held-out-model, held-out-block evaluation. Every number below uses per-held-out-model leave-one-model-out agreement features, checked to match the main evaluation engine exactly.

View ladders. The full ladders answer the budget question of Section 2.3 on the new domains. GPQA (AUROC): A .719, B .663, C .704, A+B .765, A+C .770, A+B+C .786; the A+B+C$-$A+C increment is $+.016$ $[+.001, +.030]$. MedQA: A .765, B .582, C .867, A+B .806, A+C .892, A+B+C .895; increment $+.003$ $[+.00002, +.007]$; the fusion$-$self-report AUROC gaps are $+.051$ $[+.016, +.084]$ on GPQA and $+.127$ $[+.097, +.159]$ on MedQA. TruthfulQA: A .749, B .518, C .646, A+B .753, A+C .764, A+B+C .763; increment $-.001$ $[-.005, +.003]$; the fusion$-$self-report gap is $+.016$ $[-.001, +.032]$ on AUROC, crossing zero, while the paired Brier improvement excludes zero ($-.006$ $[-.010, -.002]$), so the seventh dataset's fusion gain is resolved on the proper

score but not on the ranking. On these three panels eight target resamples add at most $+.016$ AUROC; four peer calls carry the view that matters.

The misconception boundary. On thirty TruthfulQA questions all five models give the same wrong answer. On their 150 rows, resample self-agreement averages .861, against .712 on other incorrect rows: the answers models share wrongly are the answers they hold most stably. B does not rescue the peer view's boundary; it inherits it in sharper form, which is the evidence behind the routing guidance of Section 6. Tan et al. (2025) study the same class of error within one model: consistency-based detectors fall to chance on self-consistent errors, and because such errors rarely coincide across LLMs they propose a cross-model probe; the thirty questions here are the complementary case, where five models coincide.

Cheap peers in the hard regime. Replacing the frozen panel with three roughly twenty-times-cheaper peers on GPQA (effective accuracies $.348/.389/.606$) weakens the standalone agreement view significantly (C $.644$ against $.704$ frozen, difference $-.061$ $[-.121, -.002]$), while the fusion contrast is inconclusive (A+C $.744$ against $.770$, CI crossing zero). Combined with the mixed standard-benchmark results of Appendix G, the summary is that cheap-peer value is dataset-dependent and degrades with the peer–target capability gap, consistent with the operating principle of Section 4.3.

M PROOFS

Statements. Section 4 states these results in prose; the formal versions follow, in the order they are used.

Working Model 1 (affine refinement)**.** *For $S^* \in (0,1)$, each g_e is $T_{a_e,b_e}(s) = \sigma(a_e \operatorname{logit}(s) + b_e)$ with $a_e > 0$.*

Proposition 1 (rank transfer)**.** *Under Assumption 1, thresholding S^* is equivalent to thresholding its score-conditional posterior $\mathbb{P}_e(Y = 1 \mid S^*)$. Hence the two scores have identical ROC and risk–coverage curves within environment e.*

Definition 1 (consensus-attractor model)**.** *Let $m \geq 1$, $0 \leq \alpha < 1$, $0 < \rho < 1$, $0 < \tau < 1$, and $0 < c < \pi < 1$. A question is, independently: with probability α, an* attractor *question, on which all $m + 1$ panel members produce one common semantic cluster whose answer is correct with probability τ; otherwise a* regular *question, on which the target is correct with probability ρ and, given target correctness Y, each of the m peers independently joins the target's cluster with probability π (if $Y = 1$) or c (if $Y = 0$). Let k be the number of agreeing peers.*

Proposition 2 (label-only unanimity correctness)**.** *Under Definition 1, hold (α, ρ, π, c) fixed. The distribution of the agreement count k does not depend on τ. Consequently, for hypotheses that differ only in τ, every test based on unlabeled agreement counts has power equal to its size. By contrast, n labeled anchors from the unanimity cell estimate $\mathbb{P}(Y{=}1 \mid k{=}m)$ at the parametric $n^{-1/2}$ rate.*

Proposition 1. Fix an environment e and write $g = g_e$, which is strictly increasing by Assumption 1. For any threshold t, strict monotonicity gives the event identity

$$\{S^* > t\} = \{g(S^*) > g(t)\},$$

because $s > t$ implies $g(s) > g(t)$ and $s \leq t$ implies $g(s) \leq g(t)$. As t ranges over the score's range, $g(t)$ ranges over the posterior's range, so the two families of acceptance sets, $\{S^* > t\}$ indexed by t and $\{g(S^*) > u\}$ indexed by u, coincide up to relabeling of thresholds. An ROC curve records, for each acceptance set, the pair (false-positive rate, true-positive rate); a risk–coverage curve records, for each acceptance set, its probability mass and its conditional error rate. Both are functionals of the score only through this family of events, so they are identical for S^* and for the posterior $g(S^*)$. The pairwise form is immediate: $S_i^* > S_j^*$ if and only if $g(S_i^*) > g(S_j^*)$, so the two scores induce the same ranking of any collection of answers. All statements are within the fixed environment e: for $e \neq e'$ the links g_e and $g_{e'}$ may differ, so pooling raw scores across environments on one probability scale is a separate, level-alignment question. Here the score-conditional posterior means $\mathbb{P}_e(Y{=}1 \mid S^*)$, the posterior given the score. It does not condition on all information in the environment. □

Proposition 2. τ enters the joint law of (k, Y) only through $\mathbb{P}(Y \mid \text{attractor})$. The label-free law of the agreement count, $\mathbb{P}(k) = (1-\alpha)\{\rho\,\mathrm{Bin}(m,\pi) + (1-\rho)\,\mathrm{Bin}(m,c)\}(k) + \alpha\,\mathbf{1}\{k{=}m\}$, marginalizes Y out and does not contain τ. Every statistic of the unlabeled agreement count therefore has the same distribution at each value of τ, and the rejection probability is identical under the two hypotheses. Conditional labels in the unanimity cell are Bernoulli observations with success probability $\mathbb{P}(Y{=}1 \mid k{=}m)$, whose sample mean has the $n^{-1/2}$ rate. □

A parameter can be recovered from unlabeled observations only if it changes their law, and τ leaves the agreement-count law unchanged. Under the perfectly unanimous attractor of Definition 1, labels observed below unanimity carry no information about τ because $\mathbb{P}(Y \mid k{<}m)$ does not involve it, so labels intended to estimate τ must come from unanimity; an attractor that sometimes produced disagreement would spread that information below unanimity. The statement concerns the agreement observation law. Signals outside the model, such as question text or retrieval, are the external-evidence channel of Section 6.

Offset cancellation under the affine refinement. The affine refinement makes the offset-cancellation mechanism of Section 4.1 explicit. Under Working Model 1, for scores $s_i, s_j \in (0,1)$ (implementations clip scores away from 0 and 1 before applying the logit),

$$\begin{aligned}&\operatorname{logit}\mathbb{P}_e(Y{=}1 \mid S^*{=}s_i) - \operatorname{logit}\mathbb{P}_e(Y{=}1 \mid S^*{=}s_j)\\ &\quad= \big[a_e \operatorname{logit}(s_i) + b_e\big] - \big[a_e \operatorname{logit}(s_j) + b_e\big] = a_e\big[\operatorname{logit}(s_i) - \operatorname{logit}(s_j)\big].\end{aligned}$$

The intercept b_e cancels; the slope a_e does not, but $a_e > 0$ and the strict monotonicity of the logit give $s_i > s_j \iff \mathbb{P}_e(Y{=}1 \mid S^*{=}s_i) > \mathbb{P}_e(Y{=}1 \mid S^*{=}s_j)$. Only the additive log-odds offset is removed; the positive scale remains, which is why anchor recalibration still has a slope to estimate.

Proposition 3 (anchor complexity)**.** *Under Working Model 1 and the regularity conditions stated in the proof below, maximum-likelihood recalibration of (a_e, b_e) from n labeled target anchors attains expected excess log loss $O(1/n)$ as $n \to \infty$, with parameter dimension two. Under pure label shift, and with S^* normalized to the source posterior scale, $a_e = 1$ and the intercept equals the log prior-odds ratio, so one parameter suffices (Saerens et al., 2002; Elkan, 2001).*

Proof. Write $x = \operatorname{logit}(s)$ and $\theta = (a, b)$, so the recalibration family $p_{a,b}(s) = \sigma(a \operatorname{logit}(s) + b)$ is ordinary two-parameter logistic regression $p_\theta(x) = \sigma(ax + b)$ in the covariate x. The argument uses the standard regular-parametric assumptions, stated explicitly: the affine-logit model is correctly specified (the true link lies in the family, with parameter θ^* interior); anchors are independent draws from environment e; scores are bounded away from 0 and 1 (or clipped, as in our implementation), so x is bounded; and the anchor scores vary enough that the Fisher information $I(\theta^*)$ is nonsingular. Under these conditions maximum-likelihood asymptotics give $\sqrt{n}(\widehat{\theta} - \theta^*) \xrightarrow{d} N(0, I(\theta^*)^{-1})$. The same limit holds whether the fit is unconstrained or constrained to $a > 0$, because the true a_e is an interior positive value. Let $R(\theta)$ be the population log loss. Correct specification makes θ^* its minimizer with $\nabla R(\theta^*) = 0$, and for logistic models the Hessian of R at θ^* equals $I(\theta^*)$, so a second-order expansion gives

$$R(\widehat{\theta}) - R(\theta^*) = \tfrac{1}{2}(\widehat{\theta} - \theta^*)^\top I(\theta^*)(\widehat{\theta} - \theta^*) + o_p\big(\|\widehat{\theta} - \theta^*\|^2\big).$$

The quadratic form $n(\widehat{\theta} - \theta^*)^\top I(\theta^*)(\widehat{\theta} - \theta^*)$ converges in distribution to χ^2_d with mean d, so under standard integrability conditions the expected excess log loss is $d/(2n) + o(1/n)$; with $d = 2$ this is the claimed $O(1/n)$. This standard regular-parametric expansion makes the label cost explicit: with two parameters, the leading expected excess log loss is $1/n$. The one-parameter statement requires its own correct specification: if the intercept-only subfamily is correctly specified (that is, $a_e = 1$ in truth), its estimation error is $1/(2n) + o(1/n)$; if $a_e \neq 1$, fixing $a = 1$ trades lower variance for a misspecification bias that does not vanish with n. The $O(1/n)$ rate is asymptotic, and the estimator in the statement is the maximum-likelihood estimator constrained to a compact parameter set containing θ^* in its interior, which exists on every sample, including the separated samples on which the unconstrained maximizer does not; the small-anchor curves of Appendix K use the regularized estimator with single-class fallback described there. The small-anchor observations in the main text are empirical and provide no finite-sample guarantee.

For the label-shift statement, normalize S^* to the source posterior scale, so that $\operatorname{logit}\mathbb{P}_{\text{source}}(Y{=}1 \mid S^*{=}s) = \operatorname{logit}(s)$. Pure label shift means the class-conditional score densities (or probability mass

functions, for a discrete score) are environment-invariant, $f_e(s \mid Y{=}y) = f_y(s)$, while the class prior moves from ρ_{source} to ρ_e. Bayes' rule in each environment gives the posterior odds

$$\frac{\mathbb{P}_e(Y{=}1 \mid s)}{\mathbb{P}_e(Y{=}0 \mid s)} = \frac{f_1(s)}{f_0(s)} \cdot \frac{\rho_e}{1-\rho_e}, \qquad \frac{\mathbb{P}_{\text{source}}(Y{=}1 \mid s)}{\mathbb{P}_{\text{source}}(Y{=}0 \mid s)} = \frac{f_1(s)}{f_0(s)} \cdot \frac{\rho_{\text{source}}}{1-\rho_{\text{source}}}.$$

Dividing the two and taking logarithms cancels the common likelihood ratio $f_1(s)/f_0(s)$:

$$\begin{aligned}\operatorname{logit} \mathbb{P}_e(Y{=}1 \mid s) &= \operatorname{logit} \mathbb{P}_{\text{source}}(Y{=}1 \mid s) + \operatorname{logit}(\rho_e) - \operatorname{logit}(\rho_{\text{source}}) \\ &= \operatorname{logit}(s) + \operatorname{logit}(\rho_e) - \operatorname{logit}(\rho_{\text{source}}),\end{aligned}$$

so relative to the source-posterior scale $a_e = 1$ and $b_e = \operatorname{logit}(\rho_e) - \operatorname{logit}(\rho_{\text{source}})$ (Saerens et al., 2002; Elkan, 2001). Without that normalization the statement reads: label shift adds a constant to the log-odds of the source posterior, so it changes the intercept of the link and leaves its slope unchanged. □

The remaining results analyze the consensus-attractor model of Definition 1 (Section 4.2).

Proposition 4 (link shape, unanimity, and panel limits)**.** *Under Definition 1: (i) for $k < m$ the posterior log-odds is exactly affine in k with the regular slope $\log\{\pi(1-c)/(c(1-\pi))\}$; (ii) at $k = m$, $\mathbb{P}(Y = 1 \mid k = m) = \{\alpha\tau + (1-\alpha)\rho\pi^m\}/\{\alpha + (1-\alpha)(\rho\pi^m + (1-\rho)c^m)\}$; (iii) if additionally $0 < \alpha < 1$, then as $m \to \infty$ with parameters fixed, $\mathbb{P}(Y = 1 \mid k = m) \to \tau$: enlarging the panel cannot purify the unanimity cell beyond the attractor correctness rate; (iv) attractor questions never produce $k = 0$, so the zero-agreement cell is structurally free of attractor contamination.*

Proof. Throughout, Definition 1 supplies $m \geq 1$, $0 \leq \alpha < 1$ and the stated interior probabilities; part (iii) additionally assumes $\alpha > 0$. Boundary values excluded by the definition would make some log-odds or limits below degenerate.

(i) Attractor questions produce $k = m$ almost surely, so any observation with $k < m$ comes from a regular question. Conditional on the regular component, $k \mid Y{=}1 \sim \text{Bin}(m, \pi)$ and $k \mid Y{=}0 \sim \text{Bin}(m, c)$ with likelihoods

$$\mathbb{P}(k \mid Y{=}1) = \binom{m}{k}\pi^k(1-\pi)^{m-k}, \qquad \mathbb{P}(k \mid Y{=}0) = \binom{m}{k}c^k(1-c)^{m-k}.$$

Bayes' rule gives the posterior odds

$$\frac{\mathbb{P}(Y{=}1 \mid k)}{\mathbb{P}(Y{=}0 \mid k)} = \frac{\rho}{1-\rho} \cdot \frac{\binom{m}{k}\pi^k(1-\pi)^{m-k}}{\binom{m}{k}c^k(1-c)^{m-k}} = \frac{\rho}{1-\rho}\Big(\frac{\pi}{c}\Big)^k\Big(\frac{1-\pi}{1-c}\Big)^{m-k},$$

where the binomial coefficients cancel. Taking logarithms and regrouping,

$$\operatorname{logit} \mathbb{P}(Y{=}1 \mid k) = \underbrace{\operatorname{logit}(\rho) + m \log \frac{1-\pi}{1-c}}_{\text{intercept}} + k \underbrace{\log \frac{\pi(1-c)}{c(1-\pi)}}_{L},$$

which is affine in k with slope L. Since $\pi > c$ implies $\pi/(1-\pi) > c/(1-c)$, the slope L is strictly positive.

(ii) Collect the two sources of unanimity. Attractor questions contribute $\mathbb{P}(\text{attractor}, Y{=}1, k{=}m) = \alpha\tau$ to the numerator and α to the denominator; regular questions contribute $(1-\alpha)\rho\pi^m$ and $(1-\alpha)\{\rho\pi^m + (1-\rho)c^m\}$. Dividing,

$$\mathbb{P}(Y{=}1 \mid k{=}m) = \frac{\alpha\tau + (1-\alpha)\rho\pi^m}{\alpha + (1-\alpha)\{\rho\pi^m + (1-\rho)c^m\}}.$$

In particular, holding (π, c) fixed, this expression tends to τ as $\alpha \to 1$ (the numerator tends to τ and the denominator to 1), so with τ free the unanimity purity attainable at one separation level is dense in $(0, 1)$: separation does not constrain the unanimity cell.

(iii) With $\alpha > 0$ and $\pi, c < 1$, the terms π^m and c^m vanish geometrically as $m \to \infty$, leaving $\alpha\tau/\alpha = \tau$. Both conditions are necessary: with $\alpha = 0$ the ratio instead tends to 1 (since $\pi > c$), and with $\pi = 1$ the regular mass never vanishes.

(iv) $\mathbb{P}(k{=}0 \mid \text{attractor}) = 0$, so observing $k = 0$ identifies the regular mixture component. This identifies the component only; it does not by itself make the $k = 0$ cell calibrated or its answers wrong. □

Why the no-attractor model is misspecified. The next proposition records why ignoring the attractor component biases plug-in estimates, supporting the model choice of Section 4.2: leaving the attractor component out provably inflates the fitted join rates and misspecifies the multi-peer agreement likelihood.

Proposition 5 (marginal inflation)**.** *Under Definition 1 with $0 < \alpha < 1$, the marginal join rates satisfy $\hat{\pi} > \pi$ and $\hat{c} > c$, with attractor shares $\lambda_1 = \alpha\tau/\{\alpha\tau + (1-\alpha)\rho\}$ and $\lambda_0 = \alpha(1-\tau)/\{\alpha(1-\tau) + (1-\alpha)(1-\rho)\}$. The conditional-independence plug-in slope built from $(\hat{\pi}, \hat{c})$ therefore assumes an uncapped binomial link at inflated rates, while for $m \geq 2$ the true link is affine at the uncontaminated slope below $k = m$, receives the additional attractor mass in (ii) at $k = m$, and converges to τ as the panel grows by (iii). With one peer a mixture of Bernoulli laws is again Bernoulli, so only the inflation statement remains.*

Let Z be one peer's join indicator in the full mixture and condition on $Y = 1$. By Bayes' rule the attractor share among correct answers is

$$\lambda_1 = \mathbb{P}(\text{attractor} \mid Y{=}1) = \frac{\alpha\tau}{\alpha\tau + (1-\alpha)\rho},$$

which is strictly positive under $\alpha > 0$ and $\tau > 0$. On attractor questions the peer joins with probability 1; on regular correct questions with probability π. The law of total probability gives

$$\hat{\pi} = \mathbb{P}(Z{=}1 \mid Y{=}1) = \lambda_1 \cdot 1 + (1-\lambda_1)\pi, \qquad \hat{\pi} - \pi = \lambda_1(1-\pi) > 0,$$

where strictness uses $\lambda_1 > 0$ and $\pi < 1$. The $Y = 0$ case is identical with $\lambda_0 = \alpha(1-\tau)/\{\alpha(1-\tau) + (1-\alpha)(1-\rho)\}$ and $\hat{c} - c = \lambda_0(1-c) > 0$, using $\tau < 1$, $\rho < 1$, and $c < 1$. At boundary values of the parameters the inequalities weaken to equalities. The remaining statements restate (i) and (ii) of Proposition 4. □

The proposition makes no directional claim about the plug-in evidence slope $\mathrm{logit}(\hat{\pi}) - \mathrm{logit}(\hat{c})$, which can fall below the regular-component slope L, nor about the finite-panel plug-in posterior. The misspecification statement is structural: the independence link cannot represent the attractor mass at $k = m$, whose purity converges to τ by (iii) while any independence posterior with $\hat{\pi} \neq \hat{c}$ tends to a degenerate limit.

The separation identity of Section 4.3. On regular questions each peer's join event is a coin flip with rate π under $Y = 1$ and c under $Y = 0$. Writing $L = \log\{\pi(1-c)/(c(1-\pi))\}$ for the per-peer slope of Proposition 4(i), the log-likelihood ratio of a single join is $\log(\pi/c)$ and of a single non-join is $\log\{(1-\pi)/(1-c)\}$. The mean log-likelihood ratio under $Y = 1$ minus its mean under $Y = 0$ is

$$\left[\pi \log \tfrac{\pi}{c} + (1-\pi)\log \tfrac{1-\pi}{1-c}\right] - \left[c \log \tfrac{\pi}{c} + (1-c)\log \tfrac{1-\pi}{1-c}\right] = (\pi - c)\, L,$$

and the left side is by definition $D_{\mathrm{KL}}(\pi \,\|\, c) + D_{\mathrm{KL}}(c \,\|\, \pi) = J(\pi, c)$, the symmetrized (Jeffreys) divergence between the two join laws. Hence $(\pi - c)L = J(\pi, c)$ exactly. Precisely, J is the difference between the expected log-likelihood ratio under $Y = 1$ and under $Y = 0$; it is not the mutual information $I(Y; Z)$, which additionally depends on the class prior ρ. We verified the identity numerically to machine precision on six independently sampled parameter pairs. □

Agreement counting as an optimal test. On the regular component the join indicators $Z_1, \ldots, Z_m$ are conditionally independent given Y, so the likelihood ratio of the full vector is

$$\Lambda(z_1, \ldots, z_m) = \prod_{j=1}^{m} \frac{\pi^{z_j}(1-\pi)^{1-z_j}}{c^{z_j}(1-c)^{1-z_j}} = \left(\frac{\pi}{c}\right)^{k} \left(\frac{1-\pi}{1-c}\right)^{m-k}, \qquad \log \Lambda = m \log \frac{1-\pi}{1-c} + kL,$$

which depends on the vector only through k and is strictly increasing in k because $L > 0$. By the Neyman–Pearson lemma, thresholding k is therefore the most powerful test of $Y{=}1$ against $Y{=}0$ at its achievable levels *on regular questions and within the model*; and by Bayes' rule the posterior is the logistic map of an affine function of k, so the logistic link in k used throughout is the model's own posterior rather than a convenience choice. Under the full mixture, k remains a sufficient summary of the exchangeable join vector, but the unanimity cell $k = m$ receives the attractor mass of Proposition 4(ii), so the posterior need not continue the regular-component monotone link there: with a sufficiently impure attractor (small τ, large α), $\mathbb{P}(Y{=}1 \mid k{=}m)$ can fall below $\mathbb{P}(Y{=}1 \mid k{=}m{-}1)$ (for example, $m = 4$, $\alpha = .4$, $\tau = .2$, $\rho = .5$, $\pi = .7$, $c = .15$ gives .32 against .97). This example shows explicitly that unanimity can be less reliable than near-unanimity under the full mixture.

N BOUNDARY PROBES AND ENCODER RECALIBRATION DETAILS

Attractor-fit diagnostics. Under cross-fitting, the full attractor model matches the strongest label-free empirical baseline in mean absolute AUROC error (.035 against .034), beats its no-attractor ablation in 54 of 75 folds, and improves that ablation's error (.053). On the eight-model SimpleQA panel, observed unanimity purity rises from .516 to .679 as the panel grows to seven members rather than approaching one, the behavior summarized in Section 4.2. On the fifteen 512-question environments, the jointly refit directed divergence and $J(\pi, c)$ order agreement-count AUROC at Spearman rank correlation .936 and .896, respectively.

Count sufficiency of the C view. The model generates the target-join count; the deployed C view carries three further partition statistics (largest-cluster share, cluster entropy, unique-answer fraction). A preregistered ablation on the three 512-question panels replaces the four statistics with the agreement fraction alone under the identical double-holdout protocol, after reproducing the published per-row predictions exactly. The count carries the view: pooled AUROC moves from .837/.922/.836 to .836/.924/.832 on GSM8K, TriviaQA, and MMLU-Pro; the shape statistics help only on MMLU-Pro ($+.004$ $[+.001, +.007]$, the finite-answer regime) and slightly hurt on TriviaQA ($-.002$, interval excluding zero), and the fused deltas stay within .002 with every interval crossing zero. A count-blind signal appears where the model is structurally silent: on rows where no peer joins the target ($k = 0$, where the count is constant by construction), consensus among the peers on an answer the target avoided predicts error at AUROC .569 $[.542, .598]$ pooled over 1,709 rows, a signal the largest-cluster share carries; adding the shape statistics to the fusion yields no detectable AUROC gain. On these three panels a single permutation-invariant statistic, the fraction of peers agreeing with the target, therefore beats every exactly retrained encoder on every metric by itself.

The confirmatory pass at scale. The attractor analyses were developed on the first 128-question panels; the preregistered confirmatory pass reruns the frozen tests on the 512-question panels (which extend the GSM8K and TriviaQA development questions with 384 and 381 new ones; MMLU-Pro had no development round), with attractor parameters estimated from source models only. Five of six tests pass, including all three finite-answer-space predictions: wrong-answer coincidence rises exactly in the predicted order ($c = .078/.127/.195$ for GSM8K, TriviaQA, MMLU-Pro), the mean agreement slope shrinks on the multiple-choice set, and unanimity purity drops; the pass predicted held-out unanimity purity nearly exactly (Pearson .963, median error .006). Rerun on the 384 and 381 new questions alone, with the development questions removed, the same five tests pass and the same one fails (purity prediction Pearson .966, median error .005; slope prediction Pearson .720). The one failure, target-level deviation prediction, is the first probe below. The one failing test anticipates the boundary of Section 4.3.

The four source-only probes of Section 4.3, in detail. Four distinct source-only probes, beginning with the confirmatory pass's one failing test, fail at the same term. First, the one confirmatory failure: source-only parameters predict dataset-level slopes to within .08 but cannot predict which target deviates (pooled Pearson .75 against the frozen .80 bar), while predicting unanimity purity, a target-independent quantity, almost exactly (.963). Second, the source-row plug-in separation barely tracks C-view AUROC (Spearman .38, against .96 for the corresponding target-side plug-in). This is distinct from the jointly refit D_{KL} diagnostic in the attractor-fit paragraph above, which gives .936 against agreement-count AUROC. Third, a label-free reconstruction fails on the dispersed (non-unanimous) questions. Writing $\rho_<$, $\pi_<$, $c_<$ for the target prevalence and join rates computed on that subset, the decomposition $r_< = \rho_<\pi_< + (1 - \rho_<)c_<$ holds there exactly, and the archived inversion check recovers $\pi_<$ from the other three quantities to machine precision; these subset rates are distinct from the regular-component parameters of Definition 1. The deployable arm substitutes a self-reported estimate for $\rho_<$ and a source-model estimate for $c_<$. The source estimate transfers well (median error .018), while the self-reported $\rho_<$ is overconfident precisely on this subset (worst case: reported .766 against actual .202). Substituting each component separately attributes the failure: the self-report substitution alone yields median $\pi_<$ error .105, the source-c substitution alone .035, and both together .120, so self-report bias is the dominant, though not the only, contribution. Its concentration on hard cases is consistent with reward-driven overconfidence under reinforcement learning from human feedback (RLHF) (Leng et al., 2025). Fourth, when the fusion weights are refit with target labels, the weight on the agreement view tracks the target–panel competence gap

at Pearson $-.82$; the deployable source-only fit compresses that variation 1.59-fold and loses the correlation ($-.23$).

Why a label-free staleness trigger fails. The preregistered marginal-profile trigger fails on the frontier round for a structural reason. The stale panel's marginal agreement rate is nearly unchanged from the source generation (.517 to .510) even though its separation collapses ($\pi - c$ from .278 to .044, per-peer slope L from 1.15 to 0.18), because the rise in target accuracy and the fall in separation cancel in $r = \rho\pi + (1-\rho)c$ (all three rates estimated on the same rows), an identity these rows satisfy to within 4×10^{-4}. A preregistered profile-distance diagnostic built on those marginals therefore ranks the stale panel as closer to the source profile than the refreshed panel, which restores separation to .351 and moves the marginal rate to .868; its planted-positive control failed and the diagnostic is recorded as invalidated. Panel health lives in the relationship between agreement and correctness. The marginal agreement profile alone is insufficient, which is why Section 5 pairs the deployable score with a labeled audit instead of a source-only trigger. An exploratory diagnostic shows the same mechanism away from the extreme (run after the confirmatory round and outside the preregistration): across the fifteen 512-question environments, C-view AUROC declines with the target–panel accuracy gap at Spearman $-.73$ ($p = .002$), so the frontier collapse is the endpoint of a graded relationship rather than an isolated regime.

Encoder recalibration in detail. On the 128-question pilot the fine-tuned text-encoder baselines had fitted slopes near zero or negative. The 512-question panels sharpen the picture: the encoders retain useful ordering on average within cells (mean per-cell AUROC .58–.69 across all encoder–dataset entries), while individual cells can still invert (Appendix H). A single oracle two-parameter recalibration recovers correspondingly little (post-Platt Brier $.240/.095/.208$ against constant-prediction baselines at each dataset's prevalence, $.250/.098/.224$). The corrected statement is quantitative: weak positive ordering remains within cells; this does not establish the monotone assumption, and recalibration recovers little because the transferable ordering component is small.

O THE FRONTIER UPGRADE EXPERIMENT

Three frontier targets (Claude Fable 5, Claude Opus 5, GPT-5.6 Sol) answered GPQA-Diamond under the fixed-answer protocol: 594 stage-1 requests, 572 visible answers, 22 terminal no-answers (20 from Fable), and 1,701 of 1,716 stage-2 confidence calls valid. Effective accuracies on the request denominator are $.843/.919/.949$; the analysis plan, including the primary estimand and the scale clause, was preregistered before collection.

Primary numbers. Pooled over the three targets, transferred self-report A reaches $.784$ AUROC on 567 fully scored rows (raw direct confidence .767 on the same rows); stale-panel agreement C reaches $.522$ on 572 rows and stale fusion A+C $.664$ on 567; refreshed agreement reaches .737 on 531 rows and refreshed fusion .772 on 527. The refreshed panel is frontier-dominant (two frontier peers plus Sonnet 4.5); a fit-free diagnostic using the two frontier peers alone gives $.730$ pooled on complete-visible rows, and a fail-closed sensitivity that scores peer no-answers as disagreement gives $.793$. Mixing old and frontier peers yields $.620$ on 572 rows, showing dilution. On the 531 rows where all eight peer letters are present (28 wrong answers), the fit-free agreement fractions give $.567$ for the stale panel, $.744$ for the refreshed panel, and $.651$ for the pooled seven-peer fraction. A logistic head on the two counts, trained on the other two targets and on question folds disjoint from the scored question, has median pooled AUROC .722 over 21 fold assignments (range .606–.761; the head is high-variance at 28 errors) and assigns the stale count a negative coefficient in 312 of 315 fits; it learns a negative stale weight and recovers the refreshed ranking under joint holdout. We claim no level for the separated head. The dilution is a property of the unweighted pooled fraction. Per-target transferred level errors (calibration-in-the-large, mean predicted probability minus accuracy) are $-.285/-.310/-.273$, documenting the scale drift discussed in Section 5. Question-bootstrap AUROC contrasts on common questions: refreshed C minus stale C $+.184$ $[+.081, +.300]$; refreshed fusion minus stale fusion $+.113$ $[+.019, +.202]$; refreshed fusion minus self-report $-.014$ $[-.105, +.080]$. The first two exclude zero and the third is a null: the refresh restores the agreement signal and repairs the fusion relative to its stale form, while the refreshed fusion draws level with self-report on this sample rather than exceeding it.

Anchors. Twenty pooled labeled anchors close 93.1% of the oracle-intercept Brier gap for A and 97.5% for the stale fusion, stable at 96–99% under finite-cap and Gaussian-prior refits. Raw direct confidence on this sample already sits at Brier .047, better than the anchored transferred scores (.064–.067): when no repair is trusted, the raw report remains an insurance option.

Double shift. Training on the old datasets and models and testing on frontier GPQA shifts model and dataset simultaneously: A .785, C .524, A+C .628. The registered same-domain fusion clause had two parts: A+C above .65 passed (.664), and A+C above A failed (contrast $-.118$, 95% CI $[-.243, +.005]$), which is the collapse itself. The .65 clause also fails under double shift, which is the failure-map row: under double shift, keep the transferred ranking and drop the stale panel.

Frontier arbitration. Among the 18 decidable Fable–Sol disagreements, raw confidence picks the correct side at .500; the calibrated self-report view picks it at .722. With eighteen pairs this is an illustration, reported with its size.

Second frontier round: MMLU-Pro. The same three frontier targets answered the 512 MMLU-Pro questions (1,536 requests, four without an answer) with visible accuracy .915/.908/.885, above the old panel's mean (.66) and its strongest member (.847), a moderate gap where GPQA's was wide. Five predictions were frozen before collection and all held. Under the old calibrator, stale agreement scores .622 and the stale fusion .683, no better than self-report (.691); the refreshed panel lifts agreement to .739 and the fusion to .778 (Brier .068 against .105), $+.086$ over self-report with a paired interval $[.039, .130]$; the mixed panel sits between (.689, .735). Fit-free agreement fractions tell the same story (.635 stale, .753 refreshed, .698 mixed; refresh minus stale $+.113$ $[.041, .176]$). The stale calibrator's probabilities sit .08–.15 below the targets' accuracy (calibration-in-the-large, probability units) for all three targets; a shared intercept refit on twenty labels closes a median 89% of the Brier gap to the oracle intercept. Denominators are 1,522–1,532 rows per arm. Compared with GPQA, the difference is that refresh here adds a gain over self-report rather than restoring an option, consistent with the moderate gap leaving separation to recover.

Provider-disjoint sensitivity. The refreshed panel is frontier-dominant (the two other frontier targets plus Sonnet 4.5), so for the two Anthropic targets two of its three peers share the target's provider. The registered provider-disjoint variant drops Sonnet 4.5 from the stale panel for those targets and leaves them a single cross-provider peer (GPT-5.6 Sol) in the refreshed panel. Stale agreement moves from .635 to .618 fit-free and from .622 to .605 fitted, the stale fusion from .683 to .672, so the collapse does not depend on the same-provider peer. The one-peer refreshed agreement scores .694 pooled (.604 and .659 on the two Anthropic targets; Sol's panel is already provider-disjoint at .796), a binary feature that still ranks above the stale panel; the refresh gain is smaller under this restriction ($-.059$ $[-.092, -.030]$ against the full refreshed panel), and the fitted refreshed fusion is undefined with one peer. On GPQA the same construction loses nothing: on the 531 rows with every letter present, the provider-disjoint refreshed agreement scores .753 against .744 for the full refreshed panel ($+.008$ $[-.063, +.070]$) and $+.193$ $[+.039, +.349]$ over the provider-disjoint stale panel, with Sol as the only peer of the two Anthropic targets at .744 and .812.

Other fixed scorers at the frontier. Table 4 restricts every scorer to one common row set. On each scorer's own full valid population, paired against self-report on the same rows, the picture is the same: Qwen3-32B as a verifier ranks at .554 on GPQA (567 rows) and .612 on MMLU-Pro (1,530 rows) against self-report's .784 and .691; Haiku 4.5 at .617 and .687; the released correctness model's primary adapter at .508 and .534; the retrained encoder at .459 and .454; the stale agreement count at .522 and .622. The registered prediction that these fixed source-side scorers would hold their source-panel ranking at the frontier was not met for any of them. The separation account of Section 4.3 predicts the agreement count's degradation from the target–panel gap; the same direction for the other scorers is an observation this paper does not derive, consistent with each of them being less capable than the frontier target it scores. Refreshed same-generation peers are the one source-side scorer that recovers.

The same peers asked to judge. Two frozen plans then read the frontier answers as verdicts. In the stale arm the five previous-generation panel models judged all 2,104 frontier-target rows (10,487 of 10,520 calls usable); the primary statistic is the mean over the three strongest by source-panel ac-

curacy (Sonnet 4.5, Qwen3, Maverick), on the 546 GPQA and 1,521 MMLU-Pro rows where every stale-arm scorer is valid. Registered: that this mean would sit within .05 of stale agreement FAILED, because stale verdicts retain more (.658 against .524 on GPQA, $+.132$ $[+.020, +.249]$; .682 against .619 on MMLU-Pro, $+.062$ $[-.020, +.144]$); that it would stay below self-report PASSED (.811 and .688 on the common rows; $-.152$ $[-.238, -.061]$ and $-.006$ $[-.062, +.048]$); and the paired per-judge degradation, each judge scored on the same questions for the source-panel targets and for the frontier targets, PASSED on MMLU-Pro (Sonnet 4.5 .914 $\rightarrow$ .682, Qwen3 .868 $\rightarrow$ .656, Maverick .877 $\rightarrow$ .665, every interval excluding zero) and FAILED as registered on GPQA (.893 $\rightarrow$ .668, .783 $\rightarrow$ .564, .810 $\rightarrow$ .699; Maverick's interval $[-.009, +.231]$ includes zero). In the refreshed arm the Table 3 refreshed panel judged the same rows at the answering-time reasoning budget (6,180 of 6,312 calls usable). On the 459 and 1,489 rows valid in both arms (Table 4 further requires the encoder and the correctness model, 453 and 1,431), refreshed verdicts reach .774 on GPQA and .785 on MMLU-Pro against stale verdicts .635 and .665 ($+.138$ $[-.034, +.280]$ with 20 wrong answers; $+.120$ $[+.064, +.182]$), refreshed agreement .662 and .716 (at or above it minus .02 on both, the registered clause PASSED), and self-report .789 and .677 ($-.013$ $[-.131, +.091]$; $+.108$ $[+.065, +.158]$); the registered clause that refreshed verdicts beat self-report on GPQA FAILED, and the clause that they beat stale verdicts on both datasets FAILED on the GPQA interval. Per judge on their own rows, the three frontier judges reach .78–.87 on GPQA and .70–.82 on MMLU-Pro while the panel's previous-generation member, Sonnet 4.5, sits at .65 and .67; the pooled verdict inherits that dilution (an observation outside the registered tests). Because a thinking-model judge sometimes exhausts the registered 8,000-token budget on the hardest questions, the registered row exclusion drops 103 rows that are harder than the kept ones (frontier accuracy .873 against .949 on GPQA), so a recorded amendment adds two sensitivities: a separate 16,000-token attempt recovers 57 of the 109 exhausted calls, and a fail-closed treatment scores a missing judge as disagreement and drops no row. Under both, the stale and refreshed verdict means and self-report move by at most .03 and no contrast changes sign or interval status (refreshed verdicts .775/.775/.776 on GPQA against self-report .789/.799/.788 under the three treatments; .785/.792/.799 against .677/.686/.692 on MMLU-Pro). Every registered verdict, the panel means, and the paired contrasts were recomputed by an independent implementation before use.

P DECISIONS IN DETAIL

Calibrated scores exist to make decisions. This section evaluates the two decisions the estimand was built for: choosing between two disagreeing models, and choosing which answers to accept without review.

Arbitration between disagreeing models. When two deployed models give different answers, raw confidence is a poor referee because it saturates: the strongest models report high confidence on right and wrong answers alike, and on the eighteen frontier disagreements between Fable 5 and Sol the raw reports pick the correct side exactly half the time. Across GSM8K, TriviaQA, MMLU-Pro, GPQA-Diamond, and TruthfulQA, 5,645 answer-cluster disagreements are *decidable*: exactly one model's answer is correct. Picking the answer with the higher raw confidence resolves .794 of them; picking by the calibrated A+C score resolves .829, a pooled gain of $+.036$ $[+.020, +.052]$ under question-cluster bootstrap; each score comes from a held-out fit that excludes its model. The gain is not uniform: on GSM8K raw confidence wins (.933 against .863), because there self-report is already sharp and the panel adds cross-model level noise; we report the cell rather than average it away. On the frontier pairs the calibrated self-report view flips the coin to .722; with eighteen pairs this is an illustration and carries no population-level claim. The reported A+C arm uses ten post-answer calls per pair (three confidence probes per candidate, three further panel answers, one joint clustering call), against two direct reports for the raw referee; the candidates are mutual peers and the remaining three models provide the external panel. The one-instrument deployment variant would reduce the incremental cost from eight to four calls, but is not evaluated here. Arbitration compares two models on one scale; each score comes from a fit that excludes the scored model, so cross-model commensurability is tested rather than assumed, which is what the pooled evaluation of Section 3 measures.

Routing to automation or review. The risk–coverage operating point of Section 6 carries the headline; the label economics deserve one plain statement. Sending the lowest-scoring fraction of

traffic to a human uses only the transferred ranking: under Assumption 1, the ranking itself needs zero target labels by Proposition 1. Holding a stated error ceiling is a different contract: the threshold depends on the absolute scale, so it needs either a trusted transferred link or a small labeled audit, and the coverage figures above are empirical and provide no certificate. A certificate is available with the same kind of audit: exact-binomial fixed-sequence calibration (Bates et al., 2021; Jung et al., 2025) turns a labeled target sample into a threshold whose accepted risk is at most α with confidence $1 - \delta$, at a label price several times the scale repair: our $\alpha = .05$, $\delta = .10$ demonstration draws 120 target labels, while the exact-binomial zero-error rejection requires 45 of those labels above the accepted threshold (Appendix Q). Finally, the routing rule must not treat unanimity as safety. On the thirty TruthfulQA (Lin et al., 2022b) questions where all five models give the same wrong answer, the models are more self-consistent than on their other errors (.861 against .712 resample self-agreement); no agreement signal, self or peer, flags them (Appendix L). Agreement detects instability, and a stable shared belief is exactly what it cannot detect; reducing that residual error requires capability-heterogeneous or external evidence, a class we distinguish but do not deploy.

Scoring the answer against replacing it. If peers carry correctness information, the cheapest alternative to scoring the target's answer is to answer with the peers' majority. Table 20 measures that alternative on the three 512-question panels with the four other panel models as peers. The majority beats the weakest target on every dataset and loses to the strongest on every dataset, so replacement pays only when the deployed model is not the strongest available; the fixed-answer estimand is the frame for the deployment this paper targets, a strong model whose individual answers must be scored. The majority's accuracy depends on how four-peer ties are handled, which are frequent on GSM8K, so three tie policies are shown.

Dataset	Target	Target acc.	Four-peer majority ties wrong	random tie	oracle tie	Ties
GSM8K	Sonnet 4.5	.930	.432	.476	.581	36%
GSM8K	Qwen3	.486	.523	.611	.844	35%
GSM8K	Maverick	.416	.537	.624	.851	34%
GSM8K	Mistral	.361	.574	.649	.854	30%
GSM8K	Nova Pro	.375	.564	.644	.849	31%
TriviaQA	Sonnet 4.5	.943	.890	.912	.935	7%
TriviaQA	Qwen3	.870	.900	.926	.948	6%
TriviaQA	Maverick	.880	.906	.921	.942	5%
TriviaQA	Mistral	.909	.906	.922	.945	5%
TriviaQA	Nova Pro	.850	.909	.928	.948	5%
MMLU-Pro	Sonnet 4.5	.848	.626	.675	.749	13%
MMLU-Pro	Qwen3	.613	.725	.775	.832	11%
MMLU-Pro	Maverick	.765	.648	.703	.783	14%
MMLU-Pro	Mistral	.551	.739	.783	.838	11%
MMLU-Pro	Nova Pro	.526	.741	.782	.830	10%

Table 20: **The peers' majority beats the weakest target and loses to the strongest.** Accuracy of the target's own answer against the largest semantic cluster among its four panel peers, on questions where all five panel rows are labeled and the joint partition is valid. Ties among the peers' clusters are scored as wrong, as the expected accuracy of a uniformly chosen tied cluster, or as the best tied cluster (an upper bound); the last column is the tie rate.

Q A CERTIFIED SELECTIVE-PREDICTION DEMONSTRATION

The routing discussion of Section 6 points to a certificate that the empirical coverage figures do not provide, and the stress test of Appendix J shows why one is wanted: a source-selected threshold does not reliably transfer. This appendix places the threshold in the target environment itself and demonstrates the certified recipe on the 512-question panels, using only rows already in this paper, with zero new collection. The construction keeps every assumption of the imported guarantee inspectable. A single A+C scorer per target environment is trained on source-model rows from two of the five question folds and then frozen; the target's rows on the remaining three folds (about 300 questions) form the deployment population, so no target-model row, deployment question, or anchor label ever enters training. Per draw, n labeled deployment rows are drawn uniformly as audit

anchors; the candidate thresholds are a descending quantile ladder of the non-anchor rows' scores, started where a rejection is statistically possible, so the ladder is label-free and independent of the anchor sample; at each threshold an exact binomial test of the null hypothesis that the accepted risk exceeds α is run at level δ on the anchors at or above it, and the walk stops at the first non-rejection, selecting the last rejecting threshold or abstaining if none passes, the risk-controlling calibration of Bates et al. (2021) in the fixed-sequence form of Jung et al. (2025). The imported guarantee: with probability at least $1 - \delta$ over the anchor draw, the conditional risk above the selected threshold is at most α, treating the environment's rows as draws from a common question distribution. The certificate targets that distribution. The non-anchor rows are a descriptive evaluation set, and certified coverage sits below the full-label empirical operating points of Section 6 by construction; the two must not be mixed in one comparison. We use $\delta = .10$ and 1,000 anchor draws per cell. A Monte-Carlo validity check of the walk and per-draw ladder on known data-generating processes, including a non-monotone risk curve and a heavy-ties score, keeps the violation rate of the guarantee event at or below .066 across all 27 configurations, within δ; because the scorer is fixed before calibration, those simulations match this pipeline's fixed-scorer sampling assumption.

Label economics are driven by one constant: a zero-error rejection at $\delta = .10$ needs an accepted-anchor pool of $45/22/11$ at $\alpha = .05/.10/.20$, so twenty anchors cannot certify $\alpha \leq .10$ at all, and $\alpha = .05$ is served only at $n = 120$ (TriviaQA mean unconditional certified coverage .534). At $\alpha = .10$ with $n = 120$ anchors, Table 21 gives the per-target TriviaQA results: certified coverage reaches .91–.98 conditional on certifying, which is 95–98% of the full-label oracle, with abstention near zero. At $\alpha = .20$, sixty anchors already suffice on TriviaQA (mean unconditional coverage .979, abstention .00). The harder panels behave as their base error rates dictate: GSM8K reaches mean unconditional coverage .449 at $\alpha = .10$ and $n = 120$, and MMLU-Pro .522 at $\alpha = .20$ and $n = 120$. These numbers demonstrate the label-price statement of Section 6; they are outside the headline operating points.

Target	Abstain	Coverage (certified)	Coverage (uncond.)	Oracle
Maverick	.000	.936	.936	.977
Mistral	.000	.960	.960	.997
Nova Pro	.002	.908	.906	.955
Qwen3	.000	.917	.917	.966
Sonnet 4.5	.001	.982	.981	1.000

Table 21: **With a fixed scorer, 120 anchors certify 95–98% of oracle coverage on TriviaQA with near-zero abstention.** Results use A+C with $\alpha = .10$, $\delta = .10$, and 1,000 draws per target, evaluated on the non-anchor rows. "Abstain" is the fraction of draws that certify no threshold; certified coverage conditions on certifying; unconditional coverage counts abstention as zero coverage; "Oracle" is the per-draw best coverage available with all non-anchor labels at the same risk level, averaged over draws.

R Extended related work

Surveys organize LLM confidence methods by logits, internal states, linguistic reports, consistency, and surrogate models (Geng et al., 2024). Our setting is deliberately black-box and answer-specific: methods that consume logits or hidden states, such as the LitCab bias-head calibrator (Liu et al., 2024), are excluded because commercial APIs increasingly do not expose what they need, and Xiong et al. (2024) found that logit access brought only a modest average gain in their comparison. We build on verbal self-evaluation (Lin et al., 2022a; Kadavath et al., 2022; Tian et al., 2023) and sampling-based uncertainty (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024; Lin et al., 2024; Bouchard & Chauhan, 2025). Kuhn et al. (2023) built uncertainty on a linguistic invariance; ours is statistical, asking which properties of a score survive a change of environment. Xiong et al. (2024) fuse verbalized confidence with sampling consistency within one model; our fusion crosses model boundaries, and our theory says why the two carry different information. Several works anticipate parts of the peer view: Shrivastava et al. (2023) score a closed model's answers with one open surrogate's answer probabilities, Zhang et al. (2023) use cross-model checks as an unsupervised hallucination flag, Xia et al. (2025) feed multi-model agreement into a trained text encoder, and Ning et al. (2025) optimize panel-consistency weights to rank the models themselves.

Our C view differs from all four in the same way: permutation-invariant agreement statistics about a fixed answer, fused into a probability score whose scale is audited locally, evaluated on a target model and question block the supervision never saw.

Direct comparison with AU/EU/TU. We implement the ablated $n = 2$ AU/EU/TU estimators of Hamidieh et al. (2026) under our held-out evaluation protocol, using one joint semantic partition per question in place of their similarity construction, and source-only logistic calibration for TU proper scores. Their aleatoric term (AU) is a target's intra-model response inconsistency, their epistemic term (EU) the inter-minus-intra-model similarity gap, and total uncertainty is $\mathrm{TU} = \mathrm{AU} + \mathrm{EU}$, so EU is their cross-model term, the neighbor of our C view. Their central TU-over-AU prediction replicates on all three datasets in Table 22, where each term alone ranks far below the sum. Our registered prediction that A+C would beat TU in at least fourteen of fifteen environments failed: A+C wins twelve. TU wins clearly on the GSM8K Nova Pro link-failure cell (.932 against .846) and by at most .010 in two TriviaQA cells. After source-only calibration, TU has worse Brier score than A+C in fourteen of fifteen environments. AU/EU/TU score the response distribution before an answer is fixed, so this is a neighboring uncertainty estimand; at $n = 2$, AU is also binary.

A faithful replication with their native similarity (sentence-T5-xl embeddings over each target's two resamples and the peers' answers, no partition adaptation) changes the picture in one direction only. On TriviaQA the native TU ranks at .933 and A+TU at .947 against A+C's .933, a level comparison (paired difference interval $[-0.046, +0.010]$). On GSM8K and MMLU-Pro the native TU ranks at .522 and .500, near chance, because text similarity does not separate numeric or letter answers that differ in one token; the partition adaptation used above was favorable to their method on those short-answer datasets. The three registered predictions for this replication were not met (2,502 of 2,537 MMLU-Pro rows carried two resamples; the other datasets were complete).

Is the gain the peer construction or the added self-report? A factorial run answers this on the common rows (2,553/2,537/2,502) with the published folds and head, replacing the C view by $1 - \mathrm{TU}$ while holding A fixed. Adding A to TU helps everywhere (A+TU beats A alone on all three datasets), so part of the fusion gain is self-report. With A fixed, A+C is clearly ahead on GSM8K and MMLU-Pro and level with A+TU on TriviaQA: A+TU minus A+C in pooled AUROC is $-.006$ [-.010, -.001] on GSM8K, $-.003$ [-.017, +.011] on TriviaQA, where the two peer views are nearly the same signal (C .922, TU .919), and $-.052$ [-.064, -.040] on MMLU-Pro (question-cluster bootstrap). The TU feature is computed from the joint partition of all five models' resamples, so on source rows it has seen the held-out model's responses; this joint-partition sensitivity is not a strict leave-one-model-out TU comparison. Inside the fusion, count sufficiency holds as well: A+count is within .002 AUROC of A+C on every dataset, and adding TU on top of A+C changes AUROC by at most .005.

Dataset	$1-$AU	$1-$EU	$1-$TU
GSM8K	.695	.529	.861
TriviaQA	.662	.668	.933
MMLU-Pro	.590	.603	.770

Table 22: **Total uncertainty beats aleatoric uncertainty on matched questions and held-out targets.** Pooled AUROC is reported after orienting each uncertainty as confidence; TU beats AU on all three datasets.

The closest prior transfer protocol is Pedapati et al. (2025): their logistic confidence model trains on questions answered by one LLM and transfers zero-shot to another LLM on the remaining questions. Their cross-model transfer tables report AUROC and AUARC (ECE in the in-model setting); we study calibrated probabilities under proper scores across the model shift, isolate peer agreement from target self-report, and separate transferred ordering from the environment-local probability scale. The closest supervised predecessors, APRICOT (Ulmer et al., 2024) and Calib-n (Xia et al., 2025), train text encoders to predict correctness from the question and answers; Section 3 retrains both exactly under our estimand. The agreement signal Calib-n identified is real, but encoding raw texts does not make its source-trained relationship to correctness transfer. Hamidieh et al. (2026)'s reference-size ablation, which keeps their ensemble useful up to a 32B reference model against 7–9B peers on TriviaQA, sits at the moderate-gap end of Figure 3; the frontier rounds of Section 5

extend the same curve to the gap at which agreement reaches chance. Xiao et al. (2025) attribute a correctness predictor's accuracy to its knowledge of the answer rather than to privileged self-knowledge, which is why the target's own report becomes the strongest zero-label signal once the target is the most knowledgeable model available, the regime of Section 5. Thermometer shares the transfer goal but consumes the model's last-layer features and logits, unavailable through a text-only API; its finding that dataset-level temperatures beat per-instance ones matches our treatment of the link as an environment-level object (Shen et al., 2024). Manggala et al. (2025) refine the calibration target within question–answer groups, complementary to our transfer axes. Risk-sensitive evaluations (Wang et al., 2026; Wu et al., 2026) consume the probability our estimator supplies, and Appendix Q instantiates the fixed-sequence threshold-selection guarantee of Jung et al. (2025) on our panels; their simulated-annotator confidence, agreement among personas of a single judge, is a within-model analogue of the cross-model consensus our C view measures. Three concurrent 2026 works are adjacent on single axes: Shen (2026) prove limits of verbalized confidence for risk-controlled deferral within one model, where we study which score properties survive a change of model and questions; Sun et al. (2026) compare confidence constructions by score granularity for selective prediction; and Kumaran et al. (2026) give a mechanistic account of how verbal confidence is computed, complementary to our behavioral finding that it measures commitment. Table 23 (Appendix S) places the closest methods on the deployment axes.

S DEPLOYMENT AXES OF THE CLOSEST METHODS

Method	Black-box	Supervision	Held-out model	Held-out questions	Calibrated p
Verbalized confidence	yes	optional in-domain	no	yes	partial
Sampling consistency	mostly	optional in-domain	no	yes	ranking
SAC3	yes	none	not tested	yes	no
Hamidieh et al.	yes	none	fit-free	yes	ranking
Pedapati et al.	yes	1,000 source-model	yes	yes	in-model ECE; not under model shift
Xiao et al. (GCM)	yes	8 source-model histories	yes	partial (212/512 MMLU-Pro stems seen)	yes; level-aligned zero-shot
APRICOT	yes	12,000 in-domain	no	yes	yes
Calib-n	yes	2,000 in-domain	no	yes	yes
Thermometer	no	labeled source tasks	no	yes	yes
A+C (ours)	yes	source panel; 0 target for rank	yes	yes	yes, after target-link audit

Table 23: **The released correctness model and A+C combine model holdout with calibrated probabilities; of those two, only A+C holds out every evaluation question from its supervision. The GCM MMLU adapter saw 212 of 512 stems. The methods differ in what they read (text against measurements) and in the audit A+C attaches to its scale.** For Hamidieh et al., "fit-free" means the original TU ranking score has no supervised confidence fit; our proper-score comparison adds source-only calibration. For A+C, "after target-link audit" means the ranking needs zero target labels under the monotone-link assumption, a per-environment working assumption with no label-free certificate; the labeled audit that validates the transferred scale or estimates a roughly twenty-anchor repair also checks the link (Section 4.1).

Standard tools, credited precedents, and what has no precedent. Two tools are standard and belong to no one: a logistic head over black-box features, and rank preservation under a strictly monotone recalibration, stated for verbalized confidence by Shen (2026). Four precedents are credited. Pedapati et al. (2025) observed that a logistic model over six input perturbations, trained on one LLM, transfers zero-shot to other LLMs on the same dataset; all six of their families perturb or resample the target itself, so in the taxonomy of Section 2.2 they are commitment measurements, which is consistent with their composite winning where the target's own behavior is most informative (GSM8K) and losing to peers elsewhere. Shrivastava et al. (2023), Xia et al. (2025), and Hamidieh et al. (2026) use other models as evidence: one open surrogate's probabilities, agreement text fed to an encoder, and a similarity gap under a scale-matched ensemble whose uniform weighting they condition on comparable validation risk, the assumption Section 5 tests. Xue et al. (2026) switch between self-consistency and one verifier on the target's own uncertainty; the switching variable we find is the peers' capability relative to the target. Xiao et al. (2025) train one correctness predictor on many models' histories and apply it to an unseen model; Appendix G evaluates their released model on our rows. Four elements have no precedent we know of: the frozen question–answer pair as the scored object, read by the same peers as an agreement count and as a verdict; question-block holdout nested inside every source split with several source models, leave-one-model-out peer features, and proper-score evaluation across the model shift; the agreement-evidence model that separates regular corroboration from shared attractors; and the generation-gap test, first observed on GPQA, preregistered and replicated on MMLU-Pro, and then run again with the same peers asked to judge. The propositions state exact consequences of the assumptions in Section 4, and the experiments test whether those assumptions are useful. Table 23 compares the closest methods on the deployment axes used throughout the paper.

T OPEN QUESTIONS BEYOND PREDICTION

This paper resolves the prediction target, and the measurements raise further scientific questions. Why does verbal confidence move when the answer is fixed? Which errors are shared by self-report, sampling, and peer models? When does model agreement reflect truth, and when does it reflect a common misconception? These questions motivate shared/private analysis, query-level stress sets, and controlled interventions. They do not require claiming that a black-box experiment recovers the model's inner belief; individual cases are useful examples and stress tests, while representative correctness labels are still required for probability calibration.