Candidate Selection Changes Language-Model Verifier Comparisons

Anonymous authors
Paper under double-blind review

Abstract

Candidate selection changes measured verifier gaps. On a separate set of 2,250 new questions, switching from Qwen-selected answers to uniform-all weighting widens the Qwen–Granite option-label AUROC gap by 0.097 [97.5% CI 0.069, 0.126], replicating a 0.100 change on an earlier 700-question shared candidate bank. An independent three-family MMLU holdout finds uniform-all AUROC advantages of 0.076–0.108 over model-chosen answers; a TruthfulQA extension tests the effect on a second benchmark. First-pass probability controls distinguish these selection effects from the incremental value of verification. Verifier comparisons require shared candidates and explicit policy weights.

Keywords: language model evaluation, answer verification, candidate selection, selective prediction, evaluation protocols

TL;DR: Answer-selection policy changes measured verifier discrimination and gaps between models in shared-candidate evaluations.

1 Introduction

A verifier is often evaluated on answers selected by another procedure: a model's own first choice, a sampled alternative, or an answer from a different model. Changing that procedure changes the population on which the verifier is scored. A higher AUROC can therefore reflect candidate selection even when the verifier and its scoring rule stay fixed.

We measure how much this choice changes comparisons between language-model verifiers. On a common candidate bank, changing from Qwen-selected answers to uniform-all weighting widens the Qwen–Granite option-label AUROC gap by 0.100. An evaluation on 2,250 new questions repeats the change: 0.0974 [97.5% CI 0.0686, 0.1262]. Shared questions alone are insufficient for a model comparison; the presented answers and their weights also matter.

This dependence reaches deployment decisions. In an exploratory transfer analysis, thresholds developed on uniform-all candidates raise both coverage and answered error risk when applied to a fixed chosen-answer deployment stream. We also compare verifier scores with the probabilities available from the initial option-scoring pass. These controls separate the effect of changing candidates from the additional information supplied by verification.

The study makes two contributions:

1. **A replicated change in the measured gap between verifiers.** Two shared-candidate evaluations find a Qwen–Granite option-label gap change of approximately 0.10, including a replication on 2,250 new questions. An independent three-family holdout establishes an all-answer AUROC advantage of 0.076–0.108. A collision-enrichment identity explains one component of the candidate-distribution change.
2. **Candidate selection changes threshold-transfer risk.** Transfer from uniform-all threshold development to a fixed chosen-answer stream raises both coverage and answered error risk across all nine tested predictor/model combinations, including under development refitting. First-pass baselines separate this policy dependence from incremental verifier information. TruthfulQA and task/protocol comparisons test where the selection effect extends.

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

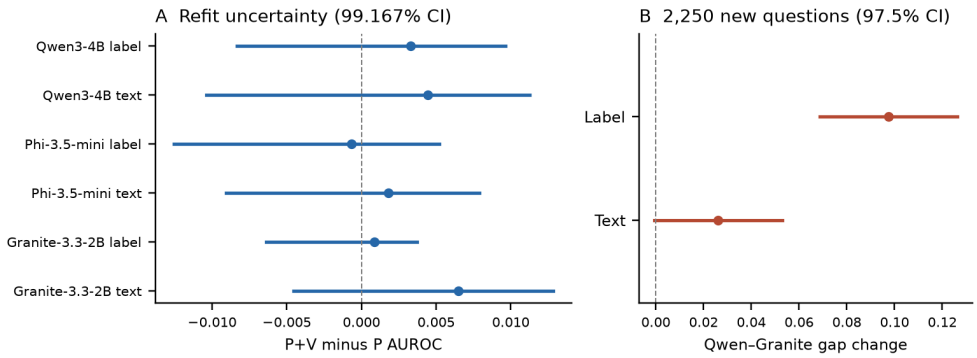

Figure 1. Shared candidates separate model comparisons from incremental verifier information. A: Incremental AUROC from adding verifier scores with joint development/evaluation bootstrap and six-comparison-adjusted 99.167% intervals. B: Verifier-gap changes on 2,250 new questions, with two-comparison-adjusted 97.5% intervals. Panel A reuses the 700-question shared bank; panel B evaluates new questions.

2 Related Work

Confidence and uncertainty. Calibration, proper scores, discrimination, and selective risk answer different questions (Guo et al., 2017; Geifman & El-Yaniv, 2017). Language-model estimators include token probabilities, verbalized confidence, and semantic uncertainty (Kadavath et al., 2022; Lin et al., 2022; Kuhn et al., 2023; Farquhar et al., 2024). ADVICE conditions on the realized answer (Seo et al., 2026); CALIBER distinguishes confidence before and after reasoning (Finlay et al., 2026). ECE also depends on binning (Chidambaram et al., 2024). We test discrimination under explicit answer-selection policies and report proper scores separately. Phalod (Phalod, 2026) compares same-model verification with likelihood baselines and selective prediction across tasks. We compare three candidate policies on a shared answer bank, quantify their effects on verifier gaps and decision thresholds, and derive a collision identity for one component of the selection effect.

Verification and self-evaluation. Revising or evaluating an answer does not itself establish access to the computation that produced it (Huang et al., 2024; Madaan et al., 2023). Self-preference in generated-text evaluation (Panickssery et al., 2024) motivates comparisons across candidate sources. Such comparisons also change candidate composition. Metacognition concerns monitoring first-order performance (Fleming & Lau, 2014); learned self-knowledge and introspection studies examine other information channels (Betley et al., 2025; Binder et al., 2025; Song et al., 2025). Our second-pass verifier sees task text and has no first-pass hidden-state channel. Verification also differs in readout granularity: outcome reward models score a realized final answer (Cobbe et al., 2021), process reward models score intermediate solution steps (Uesato et al., 2022; Lightman et al., 2023), and self-consistency aggregates multiple sampled solutions (Wang et al., 2023). This paper evaluates outcome-level readouts throughout.

Mechanistic tests. Decodability does not establish functional use (Alain & Bengio, 2017; Hewitt & Liang, 2019; Belinkov, 2022). Amnesic projection, causal abstraction, and activation steering motivate controlled interventions (Elazar et al., 2021; Geiger et al., 2021; Rimsky et al., 2024; Turner et al., 2024), whose interpretation still requires care (Jacovi & Goldberg, 2020). We test local functional dependence with projection, decision-margin and identity controls.

3 Selection Policies and Experimental Design

3.1 Multiple-choice records

For question q, gold option g, and presented option z, a forecast record is (q, z, v_z, y_z) with $y_z = \mathbf{1}[z = g]$. The first pass selects c^* by option argmax. A fresh verification pass sees the full question and presented answer and produces a normalized Yes/No score v_z. Three policies select z: **chosen** uses c^*; **uniform-all** samples from all $K = 4$ options; **uniform-nonchosen** samples from the other $K - 1$ options. The latter excludes answer-identity collisions with c^*.

We compare within-policy AUROC, defining $\Delta_{\text{all}} = A_{\text{all}} - A_{\text{chosen}}$ and $\Delta_{\text{non}} = A_{\text{non}} - A_{\text{chosen}}$. These contrasts measure discrimination under different candidate distributions. We also report Brier skill relative to each stream's own outcome base rate, defined as $1 - \text{BS}/\text{BS}_0$, where BS_0 is the Brier score of the constant forecast equal to that stream's outcome base rate.

3.2 Study design

The main comparisons use Qwen3-4B, Phi-3.5-mini, and Granite-3.3-2B. We test selection within each verifier, compare verifiers on shared candidates, and assess the predictive and decision consequences of an additional verification pass (Table 1). The independent MMLU holdout, both new-question shared-candidate evaluations and the TruthfulQA evaluation use plans specified before observing their outcomes. Predictive controls, threshold transfer, collision analyses and task/checkpoint comparisons are exploratory; they reuse previously examined questions.

Table 1. Main experimental comparisons. Full study codes, data partitions and supplementary controls are in Appendix E.

Question	Data and comparison	Design
Selection within verifiers	700 held-out MMLU questions per family; chosen, all and nonchosen answers	Independent confirmation; Section 4
Gap between verifiers	Shared candidates on 700 new questions, then 2,250 further questions	New-question comparison and replication; Section 5.1
Value beyond the first pass	100 development and 700 evaluation questions; first-pass predictors with/without verifier scores	Exploratory reuse of the shared bank; Section 5.2
Threshold transfer	Same shared bank; thresholds fitted under all-answer or chosen-answer weighting	Exploratory reuse with refitting; Section 5.3
Benchmark transfer	TruthfulQA: 684 four-option items; GSM8K: two 350-item windows	Prespecified TruthfulQA evaluation; exploratory task/protocol comparison; Section 6

The independent holdout requires complete data, at least two family point estimates $\Delta_{\text{all}} \geq 0.055$, and a three-family median ≥ 0.055. The development task/template matrix and additional checkpoint comparisons are reported in Appendices A, B and G. Appendix E gives the model coverage, item windows and threshold-calibration details.

3.3 Generated answers and direct comparisons

The GSM8K comparison uses one greedy and four sampled GSM8K completions per item (temperature 0.7, top-p 0.95, maximum 256 new tokens). The verifier sees the full proposed completion. A fixed numeric extraction and grading rule defines correctness (Appendix G). Here the third policy samples from the four **non-greedy generations**: it excludes a generation index, not necessarily the greedy answer's numeric value or meaning.

Primary and exploratory intervals use 10,000 item-bootstrap draws, pairing streams within items. For the cross-protocol and checkpoint comparisons, we average the two window-specific AUROC differences with equal weight. Task comparisons resample independently across tasks and windows; checkpoint comparisons pair items across models within each window. We report all three task comparisons with Bonferroni 98.333% intervals and all six checkpoint comparisons with Bonferroni 99.167% intervals, alongside descriptive 95% intervals. Each adjustment applies to its specified comparison set.

4 Selection Effects and Candidate Composition

4.1 The held-out selection gap

On the independent MMLU holdout, Δ_{all} is 0.1083 [0.0617, 0.1542] for Qwen3-4B, 0.0756 [0.0275, 0.1231] for Phi-3.5-mini, and 0.1044 [0.0491, 0.1593] for Granite-3.3-2B. All point estimates exceed 0.055 and the median is 0.1044, satisfying both criteria. The 12-cell development grid has positive

all-answer intervals across MMLU/ARC and standard/concise prompts, with point estimates 0.0547--0.1700. The matrix varies tasks and prompts within the same models.

All 15 small-model Δ_{non} intervals span zero. On the holdout, the three deltas are 0.0126 [-0.0592, 0.0846], -0.0217 [-0.0934, 0.0474], and 0.0002 [-0.0761, 0.0744]. The nonchosen effects are uncertain at this sample size. Appendix G tests this policy at additional checkpoints.

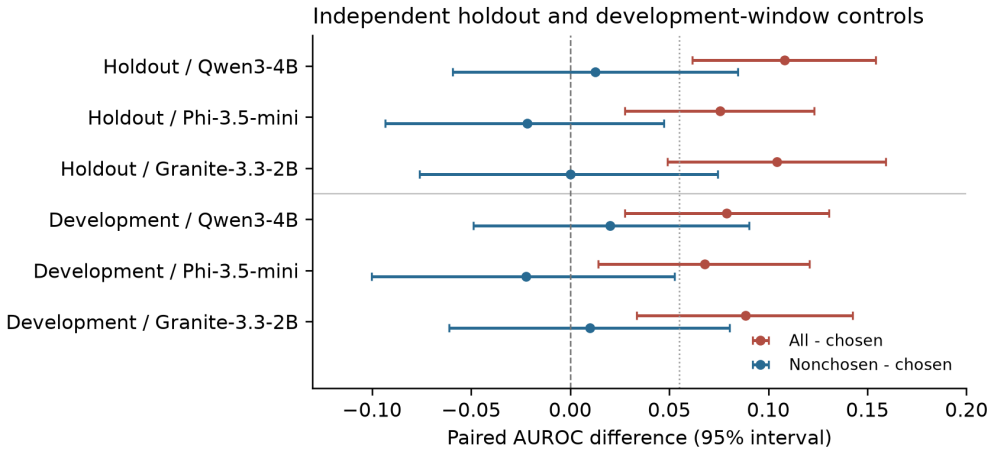

Figure 2. All-answer (red) and nonchosen (blue) AUROC contrasts in the six main cells. The 0.055 line marks the criterion applied to the holdout cells. The complete task/template matrix and within-family follow-up are in Appendices A/B.

In the Qwen2.5 follow-up, Δ_{all} is 0.0996 [0.0461, 0.1514] at 7B and 0.1328 [0.0897, 0.1770] at 14B, meeting its point-estimate rule. All 18 stream/cell combinations in the six development/holdout comparisons have negative Brier skill: discrimination gains do not imply useful probability calibration.

4.2 Why the all-answer mixture is special

Let $C = \mathbf{1}[z = c^*]$ denote a collision and let $a = P(c^* = g)$ be first-pass accuracy. With one gold option and uniform sampling independent of the first choice,

$$P(C = 1) = 1/K, \qquad P(C = 1 \mid y_z = 1) = a, \qquad P(C = 1 \mid y_z = 0) = \frac{1 - a}{K - 1}.$$

Thus whenever $a > 1/K$, collisions are enriched among correct uniform-all answers. This identity describes candidate construction. AUROC also depends on the conditional score distributions. In the six main samples, collisions occupy 0.237--0.251 of rows but 0.497--0.680 of positives.

An exploratory replacement analysis retains noncollision rows and replaces each collision row with an independently sampled nonchosen answer, its outcome, and its forecast. All six replacement-minus-chosen intervals span zero; all-minus-replacement intervals exclude zero in five cells. An exact pairwise AUROC decomposition assigns 64%--83% of uniform-all AUROC above chance to pairs involving at least one collision row (Appendix A). Replacement changes candidate identity, prevalence, and prompt content together; the decomposition attributes the observed contrast to groups of candidate pairs.

4.3 The collision indicator is an incomplete baseline

An exploratory control fits logistic predictors on 700 development items per model and evaluates them on the previously examined 700 holdout uniform-all records. Adding the clipped verifier logit to binary collision identity improves AUROC by 0.056--0.083, with positive three-comparison-adjusted intervals. The verifier adds predictive information beyond binary collision identity. Appendix I reports the fits, proper scores and intervals; Section 5 compares it with the full first-pass distribution on shared candidates.

5 Common Candidates, Predictive Value, and Decision Consequences

5.1 Comparison on new questions

The shared-candidate experiment compares verifiers on the same candidates and outcomes. For each of 700 new MMLU questions, Qwen3-4B supplies a first-pass choice and all three models score all four options under two representations: the option label or its answer text. Every verifier sees the full question and option list. All models use native BF16 on one A100 40GB. Policies assign weight one to the Qwen-selected option, $1/4$ to every option, or $1/3$ to each other option. This integrates each policy over the scored options. All verifiers evaluate the Qwen-selected option under the source-chosen policy.

The two primary estimands are the representation-specific changes in the Qwen--Granite AUROC gap, from source-chosen to uniform-all weighting. We use 10,000 bootstrap draws paired at the question level across all models, representations, and policies, with Bonferroni 97.5% intervals for these two contrasts. Each of the 700 questions contributes four candidate rows, which move together in every bootstrap draw. A variance-based rule applied to 100 disjoint development questions selected the pre-set 700-question cap before evaluation; its estimated requirement of 2,250 questions for halfwidth 0.03 exceeded that cap (Appendix J).

Table 2. Option-label and answer-text Qwen--Granite gaps under source-chosen and uniform-all weighting, with the policy change and its adjusted 97.5% interval.

Representation	Source-chosen gap	All-option gap	Change [97.5% CI]
Label	0.088	0.188	+0.100 [+0.042, +0.156]
Text	0.087	0.122	+0.034 [-0.017, +0.084]

Uniform-all weighting widens the Qwen--Granite label AUROC gap by +0.100, with the adjusted interval excluding zero. A structural asymmetry contributes to this direction: under the fixed Qwen source, the source model's collision indicator is constant, while Granite retains variance in agreement with its own first choice; uniform-all weighting restores the source model's collision variance, the enrichment quantified in Section 4.2. This interpretation is consistent with the source contrasts in Appendix K, where only the Qwen-source label interval excludes zero. The answer-text change is +0.034 with an interval spanning zero. Qwen ranks above Granite under both policies. Appendix J retains all 18 AUROC/Brier cells and the ten secondary contrasts. Granite's uniform-all label Brier is 0.573; the constant base-rate forecast scores 0.1875 (uniform-all outcomes are one gold option among four equal weights), a Brier skill of -2.05.

A replication evaluates a further 2,250 new questions, using the sample size estimated from the development data. Its two primary gap changes are +0.0974 [97.5% CI +0.0686, +0.1262] for labels and +0.0259 [-0.0007, +0.0530] for text. The label result repeats at higher precision; the adjusted text interval still crosses zero. Appendix M reports all 18 scores and 12 contrasts.

5.2 Increment beyond first-pass probabilities

The incremental-value analysis reuses the shared bank and its complete first-pass option distributions. On its 100 development questions, a logistic predictor P uses candidate logit probability, agreement with that verifier's own first choice, and first-pass entropy; $P + V$ adds the verifier logit. Features are standardized on development data, with equal question weight and fixed L2 regularization. All six evaluation comparisons use uniform-all weighting and the same 700 questions. The intervals below independently resample development and evaluation questions, refitting standardizers and predictors on every development draw; all models and readouts share the same paired draws.

Table 3. AUROC for raw candidate probability, raw verifier score, and development-fitted P and $P + V$. The last column includes development-refitting uncertainty and adjusts six comparisons. All six joint incremental intervals include zero. Raw first-pass probability has higher point AUROC than raw verification in every row. Appendix N specifies the predictor fits and joint refitting analysis; the supplement retains the conditional estimates.

Model	Readout	Raw candidate	Raw V	P	P+V	Increment [99.167% CI]
Qwen3-4B	label	0.8939	0.8559	0.8943	0.8976	+0.0033 [-0.0083, +0.0096]
Qwen3-4B	text	0.8939	0.8509	0.8943	0.8987	+0.0044 [-0.0104, +0.0113]
Phi-3.5-mini	label	0.8906	0.8140	0.8905	0.8899	-0.0007 [-0.0125, +0.0052]
Phi-3.5-mini	text	0.8906	0.8225	0.8905	0.8923	+0.0018 [-0.0091, +0.0079]
Granite-3.3-2B	label	0.7991	0.6681	0.7994	0.8002	+0.0009 [-0.0064, +0.0037]
Granite-3.3-2B	text	0.7991	0.7294	0.7994	0.8059	+0.0065 [-0.0045, +0.0128]

5.3 Candidate sources and threshold transfer

The source comparison uses each model's saved first choice to reweight the same candidate bank. The label gap changes are +0.0997, +0.0537 and +0.0210 for Qwen, Phi and Granite sources; only the Qwen-source interval excludes zero after adjustment of the six source/readout comparisons. All three adjusted text intervals include zero. These descriptive interventions jointly change the source and candidate composition (Appendix K).

The threshold-transfer analysis sets thresholds for 80% development coverage separately under uniform-all and Qwen-source-chosen candidates, then evaluates both thresholds on the same Qwen-chosen deployment candidates. Across the nine fitted predictor/model combinations, the uniform-all-derived threshold increases coverage by 0.121--0.201 and answered error risk by 0.027--0.070. The transferred threshold answers more questions and accepts a larger fraction of errors. Both contrasts retain positive pointwise 95% intervals when predictors and thresholds are refitted on resampled development questions. These intervals include variation in both question samples, conditional on the chosen predictor family (Appendices L/N).

The four-family extension adds Yi-1.5-9B-Chat on the same 700 shared-bank questions. Its label AUROC rises from 0.7179 under Qwen-chosen candidates to 0.8430 under uniform-all weighting. Qwen--Yi and Phi--Yi gap-change intervals span zero; the Granite--Yi label gap change is -0.1012 [95% CI -0.1558, -0.0469]. These are pointwise descriptive comparisons; the complete 24-contrast family appears in Appendix O.

6 Transfer Across Task Protocols and Checkpoints

6.1 Direct task-and-protocol differences

We estimate transfer directly as a difference between task-and-protocol effects. For each model, we compare the equally weighted mean Δ_{all} over its two MMLU windows with the corresponding mean over its two GSM8K windows.

Table 4. Mean Δ_{all} over two MMLU windows versus two GSM8K windows per model, with three-comparison-adjusted 98.333% intervals.

Model	MMLU mean	GSM8K mean	Difference [98.333% CI]
Qwen3-4B	+0.0936	+0.0288	+0.0648 [+0.0026, +0.1271]
Phi-3.5-mini	+0.0718	-0.0285	+0.1002 [+0.0300, +0.1721]
Granite-3.3-2B	+0.0964	+0.0075	+0.0889 [+0.0134, +0.1672]

The selection effect is larger under the MMLU protocol in all three models, with positive adjusted differences (Figure 3A). GSM8K correctness uses fixed automatic numeric extraction. A risk-stratified audit of the new-generation GSM8K cohort yields 90.48% weighted agreement [95% CI 88.20, 92.77] with 656 original human ratings; 100 overlapping ratings have three-state $\kappa = 0.601$ (Appendix P). The cross-protocol contrasts depend on grading validity and compare complete task-and-protocol packages: benchmark, candidate construction, verification prompt, precision and grading differ together. Each GSM8K effect retains its own uncertainty interval.

GSM8K chosen-answer AUROC is 0.816--0.867 for Qwen, 0.729--0.781 for Phi, and 0.346--0.393 for Granite. The three-family comparison covers both above-chance discrimination and Granite's reverse-ranking failure. The task contrasts are computed per model, not pooled across families. Excluding the reverse-ranking family leaves both remaining contrasts positive after a two-comparison adjustment (Qwen3-4B +0.0648 [+0.0053, +0.1226], Phi-3.5-mini +0.1002 [+0.0328, +0.1690]; Appendix Q). Self-consistency agreement among five completions of the same items predicts greedy correctness better than the verifier score in all three families; with two sampled completions the comparison is mixed (Appendix Q).

An extension to TruthfulQA MC1 (684 four-option items, Appendix R) replicates the selection dependence on a second outcome-level benchmark: Δ_{all} is +0.1683 [+0.1170, +0.2208] for Qwen3-4B, +0.1726 [+0.1212, +0.2258] for Phi-3.5-mini, and +0.0473 [-0.0080, +0.1024] for Granite-3.3-2B. Phi-3.5-mini's chosen-answer AUROC is 0.5040 on this benchmark while the same readout under uniform-all reaches 0.6767.

6.2 Generation indices are not answer identities

Among selected non-greedy generations, 35.7%--36.3% in Qwen, 60.9%--61.4% in Phi, and 40.3%--41.7% in Granite have the same extracted numeric answer as the greedy completion. Distinct full texts may express that shared value, and full-text identity is not retained in the released trial records. Consequently, excluding the greedy generation does not implement the multiple-choice control that excludes its answer identity. Verifier comparisons across these protocols therefore depend on both the sampling unit and the definition of an answer collision.

6.3 Additional checkpoints

Exploratory full-precision comparisons extend coverage through 32B. All eight per-window all-answer contrasts are positive, with seven 95% intervals excluding zero. The six adjusted between-checkpoint nonchosen contrasts all include zero, leaving their ordering uncertain (Figure 3B; Appendix G).

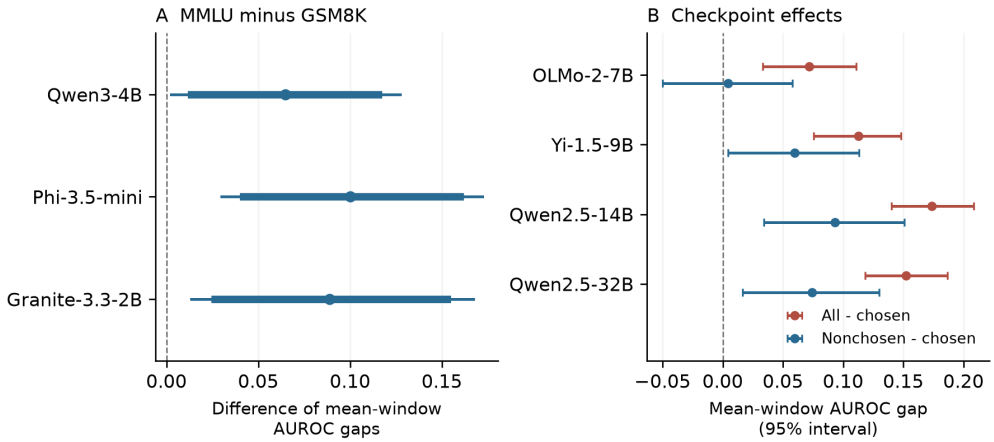

Figure 3. Exploratory task and checkpoint comparisons. A: MMLU minus GSM8K mean selection effects; thick bars show 95% intervals and thin bars the three-comparison-adjusted 98.333% intervals. B: Full-precision mean-window effects with descriptive 95% intervals. Panel B estimates effects within checkpoints; all six adjusted between-checkpoint difference intervals span zero.

7 Discussion

Candidate selection belongs in the definition of a verifier comparison. The two shared-bank evaluations hold the candidate answers fixed across verifiers and change their weights. The resulting gap changes repeat on new questions. Selection must therefore be controlled alongside model, prompt and readout when interpreting a verifier benchmark. The TruthfulQA extension adds a second multiple-

choice benchmark, while direct task/protocol contrasts and checkpoint coverage through 32B locate variation in the effect.

Deployment requires the same candidate policy used to validate thresholds. On fixed source-chosen candidates, thresholds fitted under uniform-all weighting raise coverage by 0.121–0.201 and answered error risk by 0.027–0.070. This transfer changes the operating point. Report the risk–coverage curve under the intended deployment policy, together with proper scores when forecasts are used as probabilities; the uncalibrated development and holdout streams have negative Brier skill.

First-pass scores set the relevant baseline for an additional verification pass. Raw first-pass candidate probability has higher AUROC than the raw verifier score in all six shared-bank comparisons, and all six adjusted $P+V-P$ intervals include zero. These option-scoring features are already available in this multiple-choice procedure; the tested single-pass verifier supplies no resolved increment over them. The binary collision control is weaker: adding the verifier logit improves its AUROC by 0.056–0.083. Trained verifiers and process reward models use different information and objectives (Cobbe et al., 2021; Uesato et al., 2022; Lightman et al., 2023). Self-consistency uses additional generated solutions, so its Appendix Q comparisons also carry a different sampling cost.

Question masking and single-layer projection/readout controls appear in Appendices F and C.

Limitations. The shared-candidate studies use three verifier families, one fixed source and four-choice questions. Additional sources reuse those questions. The incremental-value and threshold-transfer intervals refit predictors and thresholds under the specified features and regularization; the collision-baseline intervals condition on their development fits. Intervals containing zero leave the direction and magnitude of the effect uncertain. MMLU–GSM8K contrasts change benchmark, candidate construction, prompt, precision and grading together; these estimates use automatic numeric grading (Appendix P). Checkpoint coverage incompletely crosses scale, family and precision. The source and collision interventions change candidate composition, and the mechanistic tests cover one model/layer and one linear readout.

Conclusion. Candidate selection changes measured verifier discrimination, the gap between models and the operating point of deployed thresholds. The shared-bank gap change repeats on 2,250 new questions. Verifier evaluations should specify a shared candidate bank, its policy weights, a first-pass baseline and a threshold test on the deployment distribution.

8 Reproducibility

The anonymous artifacts contain trial records, model revisions, item windows, selection rules, seeds and CPU reanalysis for the shared-bank and exploratory comparisons. Appendices D, F and P specify the reproducibility checks, mechanistic controls and generated-answer audit. Anonymous original human labels and inclusion weights accompany the audit replay. Related anonymous submissions examine false event reports and answer/query allocation with separate fitted objects, interventions and results; their relationship is disclosed in the submission system.

AI use statement

Generative AI tools were used to critique and refine hypotheses and experimental methodology; suggest experimental parameters; implement and edit analysis and experiment code; support statistical auditing and result interpretation; generate plotting code; search for and format literature; and draft and edit manuscript text. Benchmark answer keys and fixed numeric grading define the automatic outcomes; the additional label audit uses original human ratings. All AI-assisted analyses are checked against trial-level artifacts and automated tests. The authors take responsibility for the final content and have verified this disclosure against the final workflow.

Ethics statement

This study analyzes open-weight language models and public benchmark data. It does not involve private user data or deployment on people.

Reproducibility statement

The appendix and anonymous supplement specify data partitions, model revisions, precision, seeds, intervention sites, statistical criteria, and per-trial records. The released test suite covers candidate-record construction, selection-policy implementation, calibration metrics, paired bootstrapping, the collision decomposition, and holdout separation.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop Track*, 2017.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Jan Betley, Xuchan Bao, Martín Soto, Anna Sztyber-Betley, James Chua, and Owain Evans. Tell me about yourself: LLMs are aware of their learned behaviors. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IjQ2Jtemzy.

Felix J. Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: Language models can learn about themselves by introspection. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eb5pkwIB5i.

Muthu Chidambaram, Holden Lee, Colin Mcswiggen, and Semon Rezchikov. How flawed is ECE? an analysis via logit smoothing. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8417–8435. PMLR, 2024. URL https://proceedings.mlr.press/v235/chidambaram24a.html.

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, 2021. URL https://arxiv.org/abs/2110.14168.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

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

Conor Finlay, Joshua Kurien, Saurabh Dash, Marzieh Fadaee, and Beyza Ermis. CALIBER: Calibrating confidence before and after reasoning in language models, 2026. URL https://arxiv.org/abs/2606.24281.

Stephen M. Fleming and Hakwan C. Lau. How to measure metacognition. *Frontiers in Human Neuroscience*, 8:443, 2014. doi: 10.3389/fnhum.2014.00443.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743, 2019.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=IkmD3fKBPQ.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, 2020.

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, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation. In *International Conference on Learning Representations*, 2023.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2197. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html.

Aditya Ajay Phalod. When should a language model trust itself? same-model self-verification as a conditional confidence signal, 2026. URL https://arxiv.org/abs/2605.02915.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

KiJung Seo, Sehun Lim, and Taeuk Kim. ADVICE: Answer-dependent verbalized confidence estimation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23942–23962. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1098. URL https://aclanthology.org/2026.acl-long.1098/.

Siyuan Song, Jennifer Hu, and Kyle Mahowald. Language models fail to introspect about their knowledge of language, 2025. URL https://arxiv.org/abs/2503.07513.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, et al. Solving math word problems with process- and outcome-based feedback, 2022. URL https://arxiv.org/abs/2211.14275.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

A Full C-T1R and C-T1H Results

A.1 Development and holdout cells

Cell	Chosen	All	Δ_{all} [95% CI]	Δ_{non} [95% CI]
C-T1R / Qwen3-4B	0.6688	0.7477	0.0789 [0.0277, 0.1305]	0.0202 [-0.0490, 0.0903]
C-T1R / Phi-3.5-mini	0.6769	0.7448	0.0679 [0.0141, 0.1208]	-0.0223 [-0.1001, 0.0527]
C-T1R / Granite-3.3-2B	0.6087	0.6971	0.0884 [0.0336, 0.1425]	0.0098 [-0.0610, 0.0804]
C-T1H / Qwen3-4B	0.7089	0.8171	0.1083 [0.0617, 0.1542]	0.0126 [-0.0592, 0.0846]
C-T1H / Phi-3.5-mini	0.7224	0.7981	0.0756 [0.0275, 0.1231]	-0.0217 [-0.0934, 0.0474]
C-T1H / Granite-3.3-2B	0.6030	0.7074	0.1044 [0.0491, 0.1593]	0.0002 [-0.0761, 0.0744]

The C-T1R development-cell median is 0.078906; it is reported descriptively. The C-T1H median is 0.104408, and all three holdout cells pass the preregistered family and aggregate rules.

A.2 Development task/template matrix

Model	Task	Template	all − chosen [95% CI]	nonchosen − chosen [95% CI]
Qwen3-4B	MMLU	concise	0.0547 [0.0006, 0.1074]	-0.0459 [-0.1218, 0.0307]
Qwen3-4B	ARC	standard	0.1700 [0.1148, 0.2298]	0.0531 [-0.0546, 0.1538]
Qwen3-4B	ARC	concise	0.1562 [0.1040, 0.2098]	0.0124 [-0.1033, 0.1230]
Phi-3.5-mini	MMLU	concise	0.0607 [0.0056, 0.1159]	-0.0194 [-0.0924, 0.0540]
Phi-3.5-mini	ARC	standard	0.1601 [0.0991, 0.2260]	-0.0305 [-0.1489, 0.0816]
Phi-3.5-mini	ARC	concise	0.1504 [0.0923, 0.2116]	-0.0440 [-0.1623, 0.0677]
Granite-3.3-2B	MMLU	concise	0.0684 [0.0130, 0.1213]	-0.0327 [-0.1025, 0.0376]
Granite-3.3-2B	ARC	standard	0.1274 [0.0787, 0.1763]	0.0735 [-0.0082, 0.1532]
Granite-3.3-2B	ARC	concise	0.1246 [0.0744, 0.1745]	0.0681 [-0.0135, 0.1476]

The three standard-template MMLU rows appear in Appendix A.1. The nine rows here complete the development matrix.

A.3 Brier Scores and Calibration

For each of the six development/holdout cells, chosen, `uniform-all`, and `uniform-nonchosen` Brier skill is negative. Across those 18 stream/cell combinations, values range from -1.335 to -0.054. Stream-specific ECE and Brier values are released in `paper-a-v2-selection-results.csv`; they are descriptive because policies change base rates. C-T2's 14B `uniform-all` Brier skill is 0.021.

A.4 Collision-mixture analysis

Here C denotes a collision. Replacement and removed gaps are respectively replacement-minus-chosen and all-minus-replacement AUROC, with 95% intervals.

Cell	$P(C)$	$P(C \mid +)$ / $P(C \mid -)$	Replacement gap	Removed gap
C-T1R / Qwen3-4B	0.251	0.615 / 0.136	0.002 [-0.067, 0.071]	0.077 [0.032, 0.122]
C-T1R / Phi-3.5-mini	0.237	0.621 / 0.115	-0.028 [-0.099, 0.043]	0.096 [0.056, 0.141]
C-T1R / Granite-3.3-2B	0.240	0.497 / 0.158	0.052 [-0.017, 0.119]	0.036 [-0.004, 0.078]
C-T1H / Qwen3-4B	0.239	0.675 / 0.100	0.009 [-0.061, 0.080]	0.099 [0.056, 0.146]
C-T1H / Phi-3.5-mini	0.246	0.680 / 0.107	-0.042 [-0.113, 0.027]	0.118 [0.071, 0.167]
C-T1H / Granite-3.3-2B	0.249	0.562 / 0.149	0.016 [-0.060, 0.089]	0.088 [0.043, 0.137]

The paired bootstrap resamples items, and the exact decomposition writes each collision-status positive/negative pair weight, pairwise AUROC, and chance-centered contribution. The four contributions sum to overall AUROC to machine precision. The full C-T1R task/template matrix is available in machine-readable form.

B Scale Follow-Up

Model	chosen AUROC	all AUROC	ΔAUROC [95% CI]	chosen / all Brier skill
Qwen2.5-7B	0.6552	0.7548	0.0996 [0.0461, 0.1514]	-0.884 / -0.092
Qwen2.5-14B	0.7150	0.8478	0.1328 [0.0897, 0.1770]	-1.358 / 0.021

Both point estimates and their median (0.1162) exceed the pre-specified threshold of 0.055. The two Qwen2.5-Instruct checkpoints use the same item window. C-T2 compares chosen and uniform-all candidates.

C Mechanistic Protocol and Complete Comparisons

Appendix F compares candidate sources and tests question masking on the development data. Changing the source also changes candidate identity and correctness. Masking question text while retaining the options and proposed answer reduces AUROC in all three models, showing that verification depends on the question.

Separate tests at one Qwen2.5-7B layer find an AUROC drop of 0.0779 [0.0595, 0.0972] under projection of a fitted correctness direction, with AUROC 0.0703–0.0813 below the five label-permuted projection controls. A first-pass hidden-state monitor trails the decision-margin control by 0.1039 [0.0651, 0.1437]. These tests measure local functional dependence and readout performance (Appendix C).

C.1 C-T3a directional necessity

The fit split contains 700 items and the evaluation split contains 495 correct and 205 incorrect chosen answers. A unit correctness direction is the difference between fit-split mean last-token residuals for correct and incorrect items. At the fixed `resid_post.24`, `project_out` removes each trial's component along that direction. Five controls repeat the operation with directions fitted after independent correctness-label permutations. The same 700 evaluation items are used for all streams, and every bootstrap draw is shared.

Stream relative to true-direction project-out	ΔAUROC [95% CI]
clean	0.0779 [0.0595, 0.0972]
permuted-label 0	0.0746 [0.0563, 0.0937]
permuted-label 1	0.0768 [0.0585, 0.0962]
permuted-label 2	0.0773 [0.0586, 0.0972]
permuted-label 3	0.0813 [0.0620, 0.1022]
permuted-label 4	0.0703 [0.0532, 0.0881]

Intervention energy is reported descriptively.

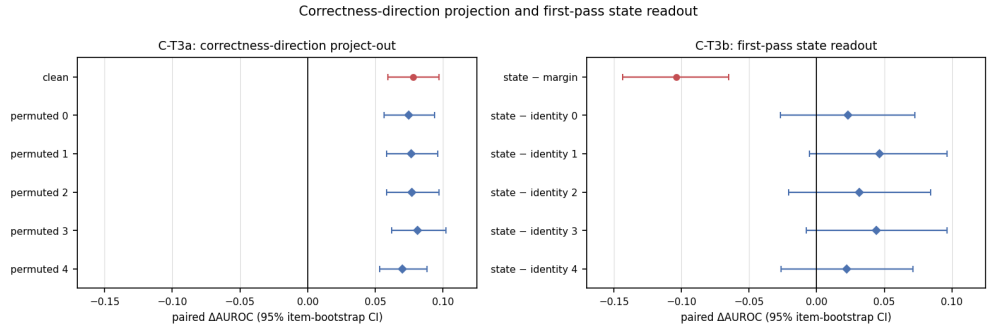

Figure C1. Left: true-direction projection relative to clean and five label-permuted controls. Right: first-pass state readout relative to decision margin and identity controls. Both show paired intervals for their separate pre-specified questions.

C.2 C-T3b first-pass state readout

State and decision-margin monitors use StandardScaler and balanced liblinear L2 logistic regression fitted on the fit split and applied unchanged to the evaluation split. The fit split has 489 correct and 211 incorrect items; the evaluation split has 473 correct and 227 incorrect items. Identity controls permute correctness labels within chosen-candidate and fit-margin-quantile bins, preserving the candidate/margin structure while destroying the intended state-label relation.

Comparison	ΔAUROC [95% CI]
state − decision margin	-0.1039 [-0.1437, -0.0651]
state − identity 0	0.0232 [-0.0270, 0.0726]
state − identity 1	0.0463 [-0.0055, 0.0962]
state − identity 2	0.0317 [-0.0208, 0.0840]
state − identity 3	0.0441 [-0.0078, 0.0964]
state − identity 4	0.0224 [-0.0264, 0.0710]

All mappings move every item while preserving their bins. The state monitor trails the decision-margin control, and all five state-minus-identity intervals include zero.

D Data Partitions and Reproducibility

MMLU test windows are disjoint across the study sequence: C-T1R `[0,700)`, C-T2 `[700,1400)`, C-T3a fit/eval `[1400,2800)`, C-T3b fit/eval `[2800,4200)`, and C-T1H `[4200,4900)`.

Independent reanalysis reproduced the study results from the specified model revisions, data windows, candidate outcomes and bootstrap samples. For C-T3, the reanalysis refits monitors, reconstructs data windows and permutation maps, and recomputes projections and bootstrap intervals from the stored trial records.

Concurrent-submission relationship. A related anonymous submission studies answer/query allocation with the same QA scorer, verifier convention, model families and public benchmarks. Its 700-question MMLU evaluation shares no normalized question stems with our A1 development, A1 evaluation or A1b evaluation banks. It shares three stems with C-T1H, two with identical choices, and one stem with the C-T3a evaluation, with different choices. Source-row windows, generated outputs, fitted objects, interventions and reported estimates remain separate. A second related submission studies fictional-event steering with separate event data and human ratings.

E Study Scope and Statistical Design

C-T1H evaluates the pre-specified all-answer gap on held-out questions in three model families. C-T1R describes variation across tasks and templates; C-T2 compares two checkpoints within Qwen2.5. C-T3a measures the effect of projecting out a fitted correctness direction relative to matched controls. C-T3b compares a state monitor with decision-margin and identity controls.

The exploratory task-and-protocol comparison estimates differences in selection effects across complete protocols. All three adjusted intervals are positive, including for Granite, whose generated-answer AUROC is below chance. Appendix Q reports the two-family comparison excluding Granite.

E7 provides exploratory full-precision coverage through 32B. Its descriptive mean-window nonchosen effects are positive for Yi and the two Qwen checkpoints, but none of the six adjusted checkpoint differences excludes zero. Between-checkpoint differences remain uncertain despite positive effects within several checkpoints.

E4 varies candidate composition together with source identity. E5 masks question text while retaining options and the candidate, measuring dependence on the question. The full-precision comparisons exclude Int4 cells. Negative Brier skill in the main cells indicates that these uncalibrated probabilities have higher squared error than the corresponding constant base-rate forecasts.

E.1 Model and scale coverage

Table E1. Model families and studies contributing selection-effect estimates.

Model	Studies
Qwen3-4B	C-T1R/H, E4/E5, E6, A0b, A1/A1b, A3/A4/A5
Phi-3.5-mini	C-T1R/H, E4/E5, E6, A0b, A1/A1b, A3/A4/A5
Granite-3.3-2B	C-T1R/H, E4/E5, E6, A0b, A1/A1b, A3/A4/A5
Qwen2.5-7B	C-T2, C-T3a/b
Qwen2.5-14B	C-T2, E7
Qwen2.5-32B	E7
OLMo-2-7B	E7
Yi-1.5-9B	E7
Yi-1.5-9B-Chat	four-family extension

E.2 Complete study index and data partitions

Study codes: C marks the candidate-selection line, E the exploratory extensions, and A the analyses of shared candidate records. C-T1R denotes the development matrix and C-T1H the independent holdout window. C-T1H, A1, A1b and T1 use evaluation plans specified before observing their outcomes. The collision analyses, A0b, A3–A5, E6b, the four-family extension and cross-protocol/checkpoint comparisons are exploratory analyses of previously examined questions. The suffixes a and b distinguish sub-studies; the codes also identify records in the accompanying data.

Table E2. Complete study index, data coverage, and designs.

Study	Coverage	Design
C-T1R	3 families, MMLU/ARC, 2 templates; 12 cells	development/descriptive
C-T1H	3 families, untouched MMLU window; 700 items each	independent confirmation
C-T2	Qwen2.5-7B/14B; 700 fresh items each	within-family all-answer follow-up
C-T3a/b	one Qwen2.5-7B layer; disjoint fit/evaluation windows	separate mechanistic tests
E4/E5	3-by-3 source cross and question masking	development on reused items
E6	GSM8K, 3 families, 2 windows of 350 items	exploratory task/protocol coverage
E7	4 full-precision checkpoints, 2 windows of 700 items	exploratory checkpoint coverage
A0b	C-T1R fits; reused C-T1H evaluation, 3 models	exploratory; coefficients fitted on development data
A1	100 disjoint development + 700 new evaluation questions, 3 models	new-question comparison
A3/A4/A5	A1 development fits, first-pass scores, source choices and thresholds	exploratory reuse of A1; refitting intervals
E6b	New generations on reused GSM8K windows	exploratory; automatic and human grading
A1b	2,250 new questions, same three verifiers and Qwen source	new-question replication
Four-family extension	Yi-1.5-9B added as a verifier on the 700 A1 questions	exploratory reuse; 24 paired contrasts
T1	TruthfulQA MC1; 684 four-option items; 3 families	new-benchmark evaluation

C-T1R uses Qwen3-4B, Phi-3.5-mini, and Granite-3.3-2B. Its standard MMLU window is $[0, 700)$; C-T1H uses $[4200, 4900)$. The pre-specified holdout criterion requires complete data, at least two family point estimates $\Delta_{\text{all}} \geq 0.055$, and a three-family median ≥ 0.055. C-T2 uses $[700, 1400)$ and requires both scales and their median to reach 0.055; it has no nonchosen control. For this two-unit C-T2 rule, 0.055 is the smallest tested threshold at or above 0.05 that meets both null-calibration criteria. The worst simulated false-pass rates are 0.00345 across empirical model pairs and 0.04685 under perfect dependence, with roughly 80% power at $N = 700$ for effects near 0.083–0.093 (Appendix E). C-T3a fits on $[1400, 2100)$ and evaluates on $[2100, 2800)$; C-T3b uses $[2800, 3500)$ and $[3500, 4200)$. E4/E5 and E7 reuse these MMLU windows. The studies span six model families from 2B to 32B; Appendix E tabulates the coverage.

E.3 C-T2 threshold calibration and power audit

We calibrate the threshold on two earlier 700-item cohorts before examining the C-T2 evaluation window. Each cohort draws 20,000 joint paired bootstrap resamples with one shared set of multinomial item weights across models; each model's distribution is centered, so no observed effect enters the null. The decision rule requires both units to reach the threshold. Candidates come from a 0.005 grid at or above the minimal relevant effect of 0.05 and pass only if the one-sided 95% Wilson upper bound on the worst null false-pass rate stays at or below 0.05 under both the empirical model pairs and a perfect-dependence stress case that copies one worst marginal-error vector to both units. The 0.055 threshold is the smallest passing candidate: 0.00345 (Wilson upper bound 0.004202) across the nine empirical model pairs, and 0.04685 (0.04937) under perfect dependence. At $N = 700$, the common true effect reaching worst-case 80% power is about 0.093 for the empirical pairs and about 0.083 under perfect dependence, with lower power at the 0.05 effect boundary. The machine-readable audit is `paper-a-v2-c-t2-power-audit.json`.

F E4/E5 Development-Study Details

F.1 Experimental setup

Both studies use the C-T1H MMLU window `[4200,4900)` and its saved source candidates, with matching gold identities for all 700 items. We use FP16 on two V100 hosts, batch-1 scoring, template `standard-v1`, seed 20260830 and the C-T1H model revisions. The design contains 12 cells $\times$ 700 rows = 8,400 trials.

Verifier	Diagonal median abs diff	Pearson r	Tolerance ($\leq$0.01 / $\geq$0.99)
qwen3-4b	0.0052	0.99739	PASS
phi35-mini	0.0038	0.99891	PASS
granite33-2b	0.0026	0.99911	PASS

F.2 E4 provenance cross

Per-cell AUROC (rows = verifier, columns = source of the presented answer; base rates 0.687 / 0.689 / 0.556 for qwen/phi/granite sources):

Verifier \ source	qwen3-4b	phi35-mini	granite33-2b
qwen3-4b	0.7100	0.7275	0.7651
phi35-mini	0.7198	0.7231	0.7701
granite33-2b	0.5927	0.5817	0.6048

Self minus pooled-other-source AUROC, 10,000 item-paired bootstrap draws with both other-source streams selected at the same resampled item indices:

Verifier	ΔAUROC [95% CI]	Interval
qwen3-4b	-0.0388 [-0.0731, -0.0055]	excludes 0
phi35-mini	-0.0216 [-0.0537, +0.0101]	crosses 0
granite33-2b	+0.0176 [-0.0192, +0.0550]	crosses 0

Source effects differ across verifiers. The Granite-source column has the highest AUROC for both other verifiers. Candidate difficulty, answer identity and base rate vary jointly across columns, while AUROC depends on the conditional score distributions.

F.3 E5 visibility ablation

Verifier	Standard	Masked	ΔAUROC [95% CI]	Band
qwen3-4b	0.7100	0.5526	+0.1574 [+0.1057, +0.2094]	large drop
phi35-mini	0.7231	0.6276	+0.0955 [+0.0492, +0.1412]	neither
granite33-2b	0.6048	0.5251	+0.0797 [+0.0255, +0.1329]	neither

The mask replaces the question text with the fixed string `[Question withheld]`; options and the presented answer are unchanged. Bands follow the pre-specified definitions (large: interval entirely above +0.05; near zero: entirely within +/-0.05). Only Qwen meets the large-drop band; Phi and Granite meet neither band, although their drop intervals both exclude zero.

G Exploratory Task and Checkpoint Comparisons

G.1 Models, grading, and analysis scope

Full-precision E7 covers OLMo-2-1124-7B-Instruct (revision 470b1fba1ae0, FP16/V100), Yi-1.5-9B-Chat (1a0fc698cf88, FP16/V100), Qwen2.5-14B-Instruct (cf98f3b3bbb4, BF16/A800), and Qwen2.5-32B-Instruct (5ede1c97bbab, BF16/A100). Both windows contain 700 MMLU items. These exploratory comparisons use full-precision checkpoints. Official GPTQ-Int4 cells are excluded because the 14B checkpoint did not meet the pre-specified numerical-insensitivity condition.

E6 uses GSM8K windows $[0, 350)$ and $[350, 700)$, the original three checkpoints in FP16, and a verifier prompted with the full generated completion. Numeric extraction takes the final match after "Answer:", falling back to the last number. Commas are removed and numeric values are compared with absolute tolerance 10^{-6}; unparsed answers count as incorrect. Extraction yield counts parsed answers across all five generations, rather than items with at least one parsed generation. Fallback extraction can select an intermediate number. Raw completions are unavailable for an independent semantic grading check. The Yes/No readout uses single-token scoring where available and continuation log probabilities for Phi.

G.2 retains the E7 run's 10,000-draw, seed-20260830 per-window intervals. G.3 and the direct comparisons use 10,000 draws and base seed 20260908. The analysis files include input hashes. The all-answer task comparison uses the C-T1R/H standard-template streams.

G.2 E7 full-precision per-window contrasts

Model/window	All minus chosen [95% CI]	Nonchosen minus chosen [95% CI]
OLMo-2-7B / dev	+0.0910 [+0.0381, +0.1449]	+0.0583 [-0.0192, +0.1338]
OLMo-2-7B / holdout	+0.0522 [-0.0025, +0.1053]	-0.0492 [-0.1261, +0.0265]
Yi-1.5-9B / dev	+0.1254 [+0.0723, +0.1762]	+0.0648 [-0.0112, +0.1410]
Yi-1.5-9B / holdout	+0.0996 [+0.0512, +0.1491]	+0.0545 [-0.0240, +0.1293]
Qwen2.5-14B / dev	+0.1906 [+0.1420, +0.2380]	+0.1038 [+0.0197, +0.1840]
Qwen2.5-14B / holdout	+0.1568 [+0.1092, +0.2058]	+0.0830 [+0.0012, +0.1630]
Qwen2.5-32B / dev	+0.1516 [+0.1019, +0.2006]	+0.0937 [+0.0153, +0.1706]
Qwen2.5-32B / holdout	+0.1524 [+0.1073, +0.2000]	+0.0544 [-0.0269, +0.1328]

All eight all-answer point estimates are positive; seven intervals exclude zero. The 14B holdout nonchosen interval is close to zero. Interval membership should not be interpreted as a discontinuous change in model behavior.

G.3 E6 per-window contrasts

"Sampled" refers to non-greedy generations, which may have the same numeric answer as the greedy generation.

Model/window	All minus greedy [95% CI]	Sampled minus greedy [95% CI]
Qwen3-4B / dev	+0.0236 [-0.0264, +0.0738]	+0.0355 [-0.0139, +0.0872]
Qwen3-4B / holdout	+0.0340 [-0.0186, +0.0870]	+0.0402 [-0.0192, +0.0986]
Phi-3.5-mini / dev	-0.0428 [-0.1070, +0.0213]	-0.0333 [-0.1029, +0.0386]
Phi-3.5-mini / holdout	-0.0141 [-0.0827, +0.0543]	-0.0338 [-0.1057, +0.0363]
Granite-3.3-2B / dev	+0.0403 [-0.0301, +0.1108]	+0.0383 [-0.0366, +0.1142]
Granite-3.3-2B / holdout	-0.0254 [-0.0948, +0.0456]	-0.0363 [-0.1176, +0.0443]

G.4 E6 operating characteristics and answer agreement

Model/window	Greedy AUROC	Greedy accuracy	Extraction yield	Numeric agreement
Qwen3-4B / dev	0.8668	0.4229	0.9954	0.3629
Qwen3-4B / holdout	0.8155	0.4429	0.9931	0.3571
Phi-3.5-mini / dev	0.7806	0.7371	1.0000	0.6086
Phi-3.5-mini / holdout	0.7293	0.6657	1.0000	0.6143
Granite-3.3-2B / dev	0.3463	0.4914	0.9994	0.4171
Granite-3.3-2B / holdout	0.3929	0.4486	0.9994	0.4029

Numeric agreement compares the selected non-greedy completion's extracted answer with the greedy answer using the same numeric tolerance; empty answers are not agreements. All extraction yields exceed 99%, satisfying the pre-specified 85% eligibility requirement. The analysis includes all three models.

G.5 Direct task-and-protocol comparisons

Let $d_{m,t,w}$ be the all-minus-chosen AUROC contrast for model m, task protocol t, and window w. We estimate $\theta_m = (d_{m,\text{MMLU},1} + d_{m,\text{MMLU},2})/2 - (d_{m,\text{GSM8K},1} + d_{m,\text{GSM8K},2})/2$. Bootstrap windows/tasks independently and retain within-item pairing across selection streams. The reported intervals are percentile intervals; adjustment allocates $0.05/3$ error to each of the three named comparisons.

Model	Difference [95% CI]	Difference [98.333% CI]
Qwen3-4B	+0.0648 [+0.0136, +0.1153]	+0.0648 [+0.0026, +0.1271]
Phi-3.5-mini	+0.1002 [+0.0420, +0.1598]	+0.1002 [+0.0300, +0.1721]
Granite-3.3-2B	+0.0889 [+0.0263, +0.1527]	+0.0889 [+0.0134, +0.1672]

The three adjusted comparisons estimate exploratory task-and-protocol differences in the selected windows. Benchmark, candidate construction, prompt, precision and grading vary together.

G.6 All six direct checkpoint comparisons

Each entry is the difference of equally weighted mean-window nonchosen effects. Items are paired across models and streams within a window after checking pool indices and gold identities; the two windows are independently resampled. Bonferroni adjustment allocates $0.05/6$ error to each comparison.

Checkpoint contrast	Difference [95% CI]	Difference [99.167% CI]
Yi-1.5-9B - OLMo-2-7B	+0.0551 [-0.0130, +0.1221]	+0.0551 [-0.0354, +0.1457]
Qwen2.5-14B - OLMo-2-7B	+0.0889 [+0.0135, +0.1641]	+0.0889 [-0.0138, +0.1903]
Qwen2.5-32B - OLMo-2-7B	+0.0695 [-0.0031, +0.1430]	+0.0695 [-0.0280, +0.1685]
Qwen2.5-14B - Yi-1.5-9B	+0.0337 [-0.0378, +0.1047]	+0.0337 [-0.0641, +0.1262]
Qwen2.5-32B - Yi-1.5-9B	+0.0144 [-0.0544, +0.0834]	+0.0144 [-0.0795, +0.1034]
Qwen2.5-32B - Qwen2.5-14B	-0.0194 [-0.0888, +0.0493]	-0.0194 [-0.1097, +0.0747]

All adjusted intervals include zero, leaving the ordering of the nonchosen effects across checkpoints uncertain.

G.7 Full-precision coverage through 32B

E7 covers OLMo-2-7B and Yi-1.5-9B in FP16 and Qwen2.5-14B/32B in BF16. All eight per-window Δ_{all} point estimates are positive; seven 95% intervals exclude zero. OLMo holdout is +0.0522 [-0.0025, +0.1053], so OLMo does not meet the pre-specified two-window replication pattern. Yi meets that pattern; the Qwen checkpoints depart from the nonchosen-null component of that pattern. The exploratory mean-window contrasts are:

Table G1. E7 mean-window all-answer and nonchosen contrasts per checkpoint with descriptive 95% intervals.

Checkpoint	All minus chosen [95% CI]	Nonchosen minus chosen [95% CI]
OLMo-2-7B	+0.0716 [+0.0332, +0.1110]	+0.0045 [-0.0499, +0.0578]
Yi-1.5-9B	+0.1125 [+0.0756, +0.1481]	+0.0597 [+0.0044, +0.1131]
Qwen2.5-14B	+0.1737 [+0.1401, +0.2086]	+0.0934 [+0.0343, +0.1507]
Qwen2.5-32B	+0.1520 [+0.1183, +0.1864]	+0.0741 [+0.0163, +0.1298]

The per-window nonchosen contrast is positive with intervals excluding zero in both 14B windows and the 32B dev window; the 32B holdout interval spans zero. The exploratory mean-window nonchosen interval is also positive for Yi, extending the observed advantage to a non-Qwen checkpoint.

All six same-item, mean-window checkpoint differences in Δ_{non} have adjusted intervals spanning zero; 32B minus 14B is -0.0194 [-0.1097, +0.0747]. These checkpoints extend the observed selection effect through 32B. Within Qwen2.5, Δ_{all} is 0.0996 at 7B and 0.1328 at 14B in C-T2, and 0.1737 and 0.1520 at 14B and 32B in E7. The direct comparisons leave its ordering across checkpoints uncertain, with family, precision and scale incompletely crossed.

H Collision-Enrichment Identity

Assume exactly one gold option per item and a candidate drawn uniformly from K options independently of the model's first-pass choice. Write $a = P(c^* = g)$, $C = \mathbf{1}[z = c^*]$, and $Y = \mathbf{1}[z = g]$. Then $P(Y = 1) = 1/K$, $P(C = 1, Y = 1) = a/K$, and $P(C = 1, Y = 0) = (1 - a)/K$. Dividing by $P(Y = 1)$ and $P(Y = 0)$ gives $P(C = 1 \mid Y = 1) = a$ and $P(C = 1 \mid Y = 0) = (1 - a)/(K - 1)$. Collisions are positively outcome-enriched precisely when $a > 1/K$.

The identity describes the sampling distribution; deterministic seeded finite samples fluctuate around it. Conditional score distributions can strengthen, erase, or reverse an AUROC gap at a given collision rate. The generated-answer protocol lacks the same single-gold-option structure and allows multiple sample indices to express the same numeric value, so the formula does not transfer by simply substituting five for K.

I Development-Fitted Predictive Controls

Model	C only	V only	C+V	Increment [98.333% CI]
Qwen3-4B	0.787	0.817	0.870	+0.083 [+0.055, +0.111]
Phi-3.5-mini	0.787	0.798	0.857	+0.070 [+0.045, +0.096]
Granite-3.3-2B	0.707	0.707	0.762	+0.056 [+0.023, +0.088]

A0b uses the standard MMLU C-T1R window $[0, 700)$ for fitting and the previously examined C-T1H window $[4200, 4900)$ for evaluation. Each model has 700 items per window. The target is correctness of the uniform-all candidate; C indicates equality with that model's first-pass choice. Let $V = \text{logit}(\text{clip}(p, 10^{-6}, 1 - 10^{-6}))$. We fit logistic predictors on C, V, (C, V), and (C, V, CV), where CV denotes the C-by-V interaction term, with an intercept, L2 regularization of inverse strength 1, the L-BFGS solver, tolerance 10^{-10}, and 2,000 maximum iterations. Evaluation labels do not enter fitting.

Model	Predictor	AUROC	Brier	Log loss
Qwen3-4B	C	0.7874	0.1252	0.4097
Qwen3-4B	V	0.8171	0.1383	0.4281
Qwen3-4B	C+V	0.8702	0.1121	0.3660
Qwen3-4B	C+V+CV	0.8703	0.1118	0.3654
Phi-3.5-mini	C	0.7866	0.1247	0.4082
Phi-3.5-mini	V	0.7981	0.1455	0.4493
Phi-3.5-mini	C+V	0.8567	0.1163	0.3803
Phi-3.5-mini	C+V+CV	0.8536	0.1146	0.3780
Granite-3.3-2B	C	0.7067	0.1536	0.4794
Granite-3.3-2B	V	0.7070	0.1650	0.5030
Granite-3.3-2B	C+V	0.7624	0.1491	0.4643
Granite-3.3-2B	C+V+CV	0.7645	0.1495	0.4649

The primary increment is additive C+V minus C-only AUROC. All models use the same 10,000 question resamples (seed 20260908), holding fitted coefficients fixed. The three exploratory model comparisons use 98.333% intervals. The interaction model and proper-score comparisons are secondary. For additive C+V versus C-only, adjusted Brier-difference intervals are [-0.0185,-0.0078], [-0.0136,-0.0030], and [-0.0103,+0.0014] for Qwen, Phi, and Granite. The corresponding log-loss intervals are [-0.0597,-0.0264], [-0.0421,-0.0132], and [-0.0319,+0.0019]. Negative differences favor the additive predictor. These predictive comparisons condition on the fitted coefficients.

J Common-Candidate Protocol and Complete Results

The question pool is MMLU test at dataset revision `c30699e8356da336a370243923dbaf21066bb9fe`. Eligible items have exactly four choices that remain distinct after NFKC and whitespace normalization. The 100 development questions and 700 evaluation questions are disjoint from each other and from the recorded earlier question-access set. A deterministic variance-based rule rounds the estimated requirement upward to a multiple of 50 and bounds it to [300,700]. Development estimated 2,250 questions for the target halfwidth 0.03; the pre-set cap selected 700. The sample size was set before evaluation; the reported intervals give the achieved precision.

Each verifier uses its named C-T1R checkpoint revision with native BF16: Qwen3-4B `1cfa9a7208912126459214e8b04321603b3df60c`, Phi-3.5-mini `2fe192450127e6a83f7441aef6e3ca586c338b77`, and Granite-3.3-2B `707f574c62054322f6b5b04b6d075f0a8f05e0f0`. The source is always Qwen3-4B. The verification prompt supplies the question and four lettered options, followed by `Proposed answer:`, either the option letter or its text, and `Is the proposed answer correct? Reply Yes or No.` The score is the normalized Yes probability. The first pass and verifier are separate calls with no transferred hidden state.

J.1 All policy-specific discrimination estimates

Verifier	Policy	Label AUROC [95% CI]	Text AUROC [95% CI]
Qwen3-4B	source chosen	0.7323 [0.6931, 0.7703]	0.7210 [0.6815, 0.7601]
Qwen3-4B	all	0.8559 [0.8363, 0.8749]	0.8509 [0.8317, 0.8695]
Qwen3-4B	nonchosen	0.7203 [0.6858, 0.7535]	0.7321 [0.6990, 0.7640]
Phi-3.5-mini	source chosen	0.6724 [0.6305, 0.7134]	0.6694 [0.6272, 0.7119]
Phi-3.5-mini	all	0.8140 [0.7940, 0.8334]	0.8225 [0.8024, 0.8417]
Phi-3.5-mini	nonchosen	0.7310 [0.6966, 0.7645]	0.7484 [0.7143, 0.7822]
Granite-3.3-2B	source chosen	0.6442 [0.5997, 0.6885]	0.6339 [0.5887, 0.6774]
Granite-3.3-2B	all	0.6681 [0.6482, 0.6876]	0.7294 [0.7082, 0.7503]
Granite-3.3-2B	nonchosen	0.5500 [0.5082, 0.5916]	0.6560 [0.6203, 0.6915]

AUROC integrates the specified candidate weights. Resampling uses 10,000 multinomial question draws (seed 20260909), paired across every model/representation/policy. All candidate rows for a resampled question share its multiplicity. The cell intervals are descriptive 95% intervals; the two primary differences of gaps have the separate 97.5% adjustment reported in Section 5.

J.2 Proper scores

Verifier	Policy	Label Brier	Text Brier
Qwen3-4B	source chosen	0.2230	0.2346
Qwen3-4B	all	0.1798	0.1689
Qwen3-4B	nonchosen	0.1654	0.1471
Phi-3.5-mini	source chosen	0.2356	0.2494
Phi-3.5-mini	all	0.2022	0.1902
Phi-3.5-mini	nonchosen	0.1911	0.1705
Granite-3.3-2B	source chosen	0.2500	0.2814
Granite-3.3-2B	all	0.5727	0.3315
Granite-3.3-2B	nonchosen	0.6803	0.3482

Brier is the question-average, policy-weighted squared error of the uncalibrated forecast. The outcome distribution varies with candidate policy.

J.3 All secondary differences of gaps

Here Q, P, and G denote Qwen, Phi, and Granite. Every row subtracts the corresponding model-pair gap under the common source-chosen policy. These ten secondary contrasts have descriptive, unadjusted 95% intervals.

Representation	Policy minus chosen	Pair	Gap change [95% CI]
label	all	Q-P	-0.0179 [-0.0579, +0.0232]
label	all	P-G	+0.1176 [+0.0724, +0.1621]
label	nonchosen	Q-G	+0.0822 [+0.0006, +0.1650]
label	nonchosen	Q-P	-0.0705 [-0.1318, -0.0082]
label	nonchosen	P-G	+0.1527 [+0.0753, +0.2297]
text	all	Q-P	-0.0232 [-0.0636, +0.0173]
text	all	P-G	+0.0576 [+0.0105, +0.1040]
text	nonchosen	Q-G	-0.0110 [-0.0772, +0.0533]
text	nonchosen	Q-P	-0.0680 [-0.1279, -0.0086]
text	nonchosen	P-G	+0.0570 [-0.0155, +0.1280]

K Candidate-Source Comparison

A4 substitutes the three saved source choices on the same 700 questions, using their existing verifier scores and candidates. For each verifier/readout, there is one shared uniform-all score and three source-specific chosen/nonchosen pairs, giving 42 distinct scoring cells. The uniform-all scores are shared across sources. All three source-specific Qwen--Granite uniform-all-minus-chosen gap changes are reported for both readouts, with a six-comparison adjustment. This exploratory analysis uses a separate bootstrap seed from A1. Its Qwen-source point contrasts estimate the same quantities, with intervals adjusted for the six source/readout comparisons.

Table K1. Changes in AUROC gaps across candidate sources. Source and candidate composition vary together. All 42 AUROC, Brier and log-loss summaries, including marginal intervals, accompany the paper in machine-readable form.

Candidate source	Readout	Qwen-Granite gap change [99.167% CI]
Qwen3-4B	label	+0.0997 [+0.0339, +0.1672]
Qwen3-4B	text	+0.0344 [-0.0231, +0.0930]
Phi-3.5-mini	label	+0.0537 [-0.0107, +0.1179]
Phi-3.5-mini	text	-0.0183 [-0.0701, +0.0349]
Granite-3.3-2B	label	+0.0210 [-0.0390, +0.0813]
Granite-3.3-2B	text	-0.0476 [-0.0999, +0.0033]

L Coverage-Threshold Transfer

These estimates condition on the development fits and selected thresholds. Appendix N gives the complementary joint refitting analysis.

A5 uses each first-pass fit and each label/text augmented fit from A3. Separately for uniform-all and Qwen-source-chosen development distributions, the threshold is the score at which descending cumulative candidate weight first reaches 80% of total development weight; all ties are accepted. Both thresholds are then evaluated on the identical Qwen-chosen deployment candidates. The scientific JSON records every threshold, achieved development coverage, deployment coverage and risk. Thresholds target development coverage only.

Table L1. Uniform-all-derived minus source-chosen-derived threshold outcomes, on the same deployment candidates. Risk is the error fraction among accepted answers; coverage is the fraction accepted. All nine combinations are retained. Paired 95% intervals are marginal and condition on fitted predictors and the development-selected thresholds. Appendix N also includes uncertainty from the 100-question development sample. The transferred thresholds accept more answers at a higher error rate.

Verifier	Predictor	Coverage change [95% CI]	Risk change [95% CI]
Qwen3-4B	P	+0.2014 [+0.1714, +0.2314]	+0.0703 [+0.0499, +0.0913]
Qwen3-4B	P+V label	+0.1514 [+0.1257, +0.1771]	+0.0542 [+0.0369, +0.0722]
Qwen3-4B	P+V text	+0.1486 [+0.1229, +0.1757]	+0.0482 [+0.0313, +0.0661]
Phi-3.5-mini	P	+0.1214 [+0.0985, +0.1471]	+0.0461 [+0.0304, +0.0629]
Phi-3.5-mini	P+V label	+0.1486 [+0.1229, +0.1757]	+0.0493 [+0.0320, +0.0670]
Phi-3.5-mini	P+V text	+0.1329 [+0.1086, +0.1586]	+0.0463 [+0.0302, +0.0634]
Granite-3.3-2B	P	+0.1657 [+0.1386, +0.1929]	+0.0275 [+0.0114, +0.0447]
Granite-3.3-2B	P+V label	+0.1671 [+0.1400, +0.1943]	+0.0341 [+0.0172, +0.0514]
Granite-3.3-2B	P+V text	+0.1957 [+0.1671, +0.2257]	+0.0432 [+0.0246, +0.0625]

M The 2,250-question shared-candidate evaluation

A1b evaluates 2,250 new MMLU questions with the A1 candidate and representation definitions. Qwen supplies the common source choice; three verifiers each score all four options. The two representation-specific Qwen--Granite gap changes are primary. The protocol was fixed before the first evaluation record. The sample size follows the earlier development precision estimate. Inference uses 10,000 paired question bootstrap draws; the two primary contrasts receive 97.5% intervals. The 9,000 option rows per verifier are clustered within 2,250 questions.

Table M1. A1b per-verifier AUROC and Brier under all three policies and both representations on the 2,250-question pool.

Readout	Policy	Verifier	AUROC [95% CI]	Brier
label	source_chosen	Qwen3-4B	0.7287 [0.7073, 0.7499]	0.2436
label	source_chosen	Phi-3.5-mini	0.7076 [0.6851, 0.7297]	0.2265
label	source_chosen	Granite-3.3-2B	0.6529 [0.6286, 0.6770]	0.2824
label	all	Qwen3-4B	0.8485 [0.8375, 0.8593]	0.1877
label	all	Phi-3.5-mini	0.8151 [0.8040, 0.8260]	0.2035
label	all	Granite-3.3-2B	0.6752 [0.6641, 0.6861]	0.5745
label	nonchosen	Qwen3-4B	0.7319 [0.7145, 0.7490]	0.1690
label	nonchosen	Phi-3.5-mini	0.7364 [0.7183, 0.7543]	0.1959
label	nonchosen	Granite-3.3-2B	0.5674 [0.5458, 0.5880]	0.6719
text	source_chosen	Qwen3-4B	0.7342 [0.7130, 0.7549]	0.2397
text	source_chosen	Phi-3.5-mini	0.7036 [0.6804, 0.7263]	0.2326
text	source_chosen	Granite-3.3-2B	0.6433 [0.6191, 0.6676]	0.2840
text	all	Qwen3-4B	0.8464 [0.8353, 0.8570]	0.1737
text	all	Phi-3.5-mini	0.8267 [0.8154, 0.8375]	0.1879
text	all	Granite-3.3-2B	0.7295 [0.7171, 0.7414]	0.3367
text	nonchosen	Qwen3-4B	0.7321 [0.7142, 0.7494]	0.1517
text	nonchosen	Phi-3.5-mini	0.7544 [0.7365, 0.7719]	0.1730
text	nonchosen	Granite-3.3-2B	0.6598 [0.6405, 0.6784]	0.3543

All policy-minus-source-chosen changes in between-verifier AUROC gaps follow. Secondary intervals are pointwise.

Table M2. All twelve policy-minus-source-chosen gap changes with pointwise intervals; the two Qwen--Granite all-option rows are the adjusted primary contrasts.

Readout	Policy	Verifier pair	Change	Interval	Level
label	all	Qwen3-4B / Granite-3.3-2B	0.0974	[0.0686, 0.1262]	97.5% primary
label	all	Qwen3-4B / Phi-3.5-mini	0.0122	[-0.0094, 0.0347]	95% descriptive
label	all	Phi-3.5-mini / Granite-3.3-2B	0.0852	[0.0610, 0.1092]	95% descriptive
label	nonchosen	Qwen3-4B / Granite-3.3-2B	0.0887	[0.0468, 0.1302]	95% descriptive
label	nonchosen	Qwen3-4B / Phi-3.5-mini	-0.0256	[-0.0575, 0.0065]	95% descriptive
label	nonchosen	Phi-3.5-mini / Granite-3.3-2B	0.1143	[0.0749, 0.1538]	95% descriptive
text	all	Qwen3-4B / Granite-3.3-2B	0.0259	[-0.0007, 0.0530]	97.5% primary
text	all	Qwen3-4B / Phi-3.5-mini	-0.0110	[-0.0315, 0.0096]	95% descriptive
text	all	Phi-3.5-mini / Granite-3.3-2B	0.0369	[0.0126, 0.0616]	95% descriptive
text	nonchosen	Qwen3-4B / Granite-3.3-2B	-0.0186	[-0.0531, 0.0163]	95% descriptive
text	nonchosen	Qwen3-4B / Phi-3.5-mini	-0.0529	[-0.0825, -0.0234]	95% descriptive
text	nonchosen	Phi-3.5-mini / Granite-3.3-2B	0.0343	[-0.0027, 0.0713]	95% descriptive

N Development-refitting uncertainty

A3 uses the 100 A1 development questions and reuses its 700 evaluation questions. For each verifier, the first-pass features are logit(p_z), $\mathbf{1}[z = c^*]$ using that verifier's own choice, and normalized first-pass entropy. Adding V means adding the corresponding label/text verifier logit. Probabilities are clipped to $[10^{-6}, 1 - 10^{-6}]$ before logits and log loss. All four candidates have weight $1/4$, so each development question has total weight one. Features are standardized with development means/scales; logistic fits use an intercept, L2 inverse regularization 1, L-BFGS, tolerance 10^{-10}, and a 2,000-iteration cap. The same uniform-all fits are evaluated under all three Qwen-source policies.

Each of 10,000 draws independently resamples development and evaluation questions, shares those draws across models and readouts, refits feature standardizers and nine logistic predictors, and reselects the 80% coverage thresholds. There are 90,009 fits including the original nine. The joint percentile intervals account for variation in both question sets, conditional on the fixed model scores, feature specification, regularization and prompt families. Development-only dispersions hold evaluation questions fixed. The six AUROC increments alone receive Bonferroni 99.167% intervals; proper scores and threshold contrasts remain pointwise. All 90 endpoints, invalid-draw counts, and both uncertainty definitions are included in the accompanying scientific data.

Verifier	Readout	Increment	Joint 95% CI	Joint 99.167% CI
Granite-3.3-2B	label	0.0009	[-0.0036, 0.0027]	[-0.0064, 0.0037]
Granite-3.3-2B	text	0.0065	[-0.0008, 0.0109]	[-0.0045, 0.0128]
Phi-3.5-mini	label	-0.0007	[-0.0091, 0.0036]	[-0.0125, 0.0052]
Phi-3.5-mini	text	0.0018	[-0.0061, 0.0063]	[-0.0091, 0.0079]
Qwen3-4B	label	0.0033	[-0.0045, 0.0079]	[-0.0083, 0.0096]
Qwen3-4B	text	0.0044	[-0.0051, 0.0094]	[-0.0104, 0.0113]

All six joint intervals include zero. Threshold transfer has positive pointwise coverage and risk contrasts in all nine predictor/model combinations:

Verifier	Predictor	Outcome	Change [joint 95% CI]
Granite-3.3-2B	augmented label	coverage	0.1671 [0.1071, 0.2371]
Granite-3.3-2B	augmented label	risk	0.0341 [0.0100, 0.0608]
Granite-3.3-2B	augmented text	coverage	0.1957 [0.0943, 0.2700]
Granite-3.3-2B	augmented text	risk	0.0432 [0.0110, 0.0731]
Granite-3.3-2B	first pass	coverage	0.1657 [0.1029, 0.2414]
Granite-3.3-2B	first pass	risk	0.0275 [0.0092, 0.0595]
Phi-3.5-mini	augmented label	coverage	0.1486 [0.0657, 0.2286]
Phi-3.5-mini	augmented label	risk	0.0493 [0.0107, 0.0853]
Phi-3.5-mini	augmented text	coverage	0.1329 [0.0729, 0.2329]
Phi-3.5-mini	augmented text	risk	0.0463 [0.0158, 0.0863]
Phi-3.5-mini	first pass	coverage	0.1214 [0.0757, 0.2457]
Phi-3.5-mini	first pass	risk	0.0461 [0.0222, 0.1012]
Qwen3-4B	augmented label	coverage	0.1514 [0.0814, 0.2629]
Qwen3-4B	augmented label	risk	0.0542 [0.0273, 0.0930]
Qwen3-4B	augmented text	coverage	0.1486 [0.0943, 0.2671]
Qwen3-4B	augmented text	risk	0.0482 [0.0293, 0.0825]
Qwen3-4B	first pass	coverage	0.2014 [0.1257, 0.2986]
Qwen3-4B	first pass	risk	0.0703 [0.0388, 0.1023]

O Four-family shared-candidate extension

Yi-1.5-9B-Chat supplies an additional BF16 verifier on the same A1 development and evaluation questions. The source choices and candidate weights are unchanged. This exploratory extension reports all six verifier pairs, two representations and two policy contrasts, with 24 pointwise intervals. Model size, training and family vary together.

Readout	Policy	Verifier pair	Gap change [95% CI]
label	all	Qwen3-4B / Phi-3.5-mini	-0.0179 [-0.0579, 0.0232]
label	all	Qwen3-4B / Granite-3.3-2B	0.0997 [0.0499, 0.1491]
label	all	Qwen3-4B / Yi-1.5-9B	-0.0015 [-0.0439, 0.0409]
label	all	Phi-3.5-mini / Granite-3.3-2B	0.1176 [0.0724, 0.1621]
label	all	Phi-3.5-mini / Yi-1.5-9B	0.0164 [-0.0274, 0.0604]
label	all	Granite-3.3-2B / Yi-1.5-9B	-0.1012 [-0.1558, -0.0469]
label	nonchosen	Qwen3-4B / Phi-3.5-mini	-0.0705 [-0.1318, -0.0082]
label	nonchosen	Qwen3-4B / Granite-3.3-2B	0.0822 [0.0006, 0.1650]
label	nonchosen	Qwen3-4B / Yi-1.5-9B	-0.0588 [-0.1187, 0.0023]
label	nonchosen	Phi-3.5-mini / Granite-3.3-2B	0.1527 [0.0753, 0.2297]
label	nonchosen	Phi-3.5-mini / Yi-1.5-9B	0.0118 [-0.0535, 0.0777]
label	nonchosen	Granite-3.3-2B / Yi-1.5-9B	-0.1409 [-0.2324, -0.0512]
text	all	Qwen3-4B / Phi-3.5-mini	-0.0232 [-0.0636, 0.0173]
text	all	Qwen3-4B / Granite-3.3-2B	0.0344 [-0.0099, 0.0776]
text	all	Qwen3-4B / Yi-1.5-9B	0.0072 [-0.0336, 0.0477]
text	all	Phi-3.5-mini / Granite-3.3-2B	0.0576 [0.0105, 0.1040]
text	all	Phi-3.5-mini / Yi-1.5-9B	0.0304 [-0.0120, 0.0721]
text	all	Granite-3.3-2B / Yi-1.5-9B	-0.0272 [-0.0755, 0.0212]
text	nonchosen	Qwen3-4B / Phi-3.5-mini	-0.0680 [-0.1279, -0.0086]
text	nonchosen	Qwen3-4B / Granite-3.3-2B	-0.0110 [-0.0772, 0.0533]
text	nonchosen	Qwen3-4B / Yi-1.5-9B	-0.0339 [-0.0932, 0.0244]
text	nonchosen	Phi-3.5-mini / Granite-3.3-2B	0.0570 [-0.0155, 0.1280]
text	nonchosen	Phi-3.5-mini / Yi-1.5-9B	0.0341 [-0.0289, 0.0964]
text	nonchosen	Granite-3.3-2B / Yi-1.5-9B	-0.0229 [-0.0955, 0.0485]

P New-generation numeric-answer audit

E6b generates five answers per question on each of the reused GSM8K development and holdout windows of 350 questions, for three checkpoints and 10,500 completions. The numeric-answer rule supplies automated outcomes. This exploratory cohort uses new generations on previously examined questions. The table reports both windows and all policy contrasts. Intervals use 10,000 paired question bootstrap draws within a window, conditional on the checkpoint and generation recipe.

Checkpoint	Window	Parse rate	Chosen AUROC	All − chosen [95% CI]	Nonchosen − chosen [95% CI]
Qwen3-4B	dev	0.9954	0.8668	0.0236 [-0.0248, 0.0750]	0.0355 [-0.0133, 0.0876]
Qwen3-4B	holdout	0.9931	0.8155	0.0340 [-0.0178, 0.0871]	0.0402 [-0.0183, 0.0992]
Phi-3.5-mini	dev	1.0000	0.7730	-0.0140 [-0.0761, 0.0501]	-0.0325 [-0.1037, 0.0376]
Phi-3.5-mini	holdout	1.0000	0.7309	-0.0021 [-0.0699, 0.0666]	-0.0191 [-0.0867, 0.0479]
Granite-3.3-2B	dev	0.9994	0.3463	0.0403 [-0.0291, 0.1105]	0.0383 [-0.0389, 0.1143]
Granite-3.3-2B	holdout	0.9994	0.3929	-0.0254 [-0.0941, 0.0434]	-0.0363 [-0.1144, 0.0433]

All twelve contrast intervals include zero. The accompanying data include each stream's AUROC, accuracy and bootstrap sample count.

P.1 Risk-stratified human label audit

The audit covers a finite pool of 3,000 completions: 200 questions, three generators and five completions per question. The questions were sampled as 100 from each original window. R1 supplies 656 ratings, sampled by numeric-extraction risk and generator; R2 supplies a nested overlap of 100. The no-extraction and marked-format-risk strata are censused, while unmarked-final-number and clean-integer strata are sampled. Each R1 rating receives the inverse of its response inclusion probability. R2 weights use the product of that probability and the conditional overlap inclusion probability. Original rating sheets remain separate, and U is retained as an indeterminate state.

R1's design-weighted agreement with the automatic labels is 90.48% [95% CI 88.20, 92.77], with 3.93% [2.28, 5.59] indeterminate. Automatic-correct/human-incorrect and automatic-incorrect/human-correct responses account for 4.27% [2.91, 5.64] and 1.31% [0.38, 2.23] of the pool. These pointwise intervals use stratified finite-population variance and retain U in every denominator. Adding the original question-sampling stage (100 of 350 questions per window) gives 87.92–93.05% for agreement and 2.10–5.77% for U, targeting the 10,500-completion bank. The two-phase calculation preserves shared question identity across generators and candidates, dev/holdout strata, and both sampling fractions.

Across the 100 overlapping responses, raw three-state agreement is 79.0% and Cohen's $\kappa = 0.601$. These overlap statistics remain descriptive: R2 has one observation in the ten-response no-extraction stratum, preventing a design-based variance estimate for that stratum. The released replay reports R1 sampling intervals and retains the original R2 point estimates. Only two question/generator groups have all five R1 labels, so this audit measures grading agreement without supplying human-label AUROC estimates. The automatic-outcome policy contrasts above retain their original definitions. Anonymous labels, question clusters, sampling-cell counts and interval replay are released in `paper-a-human-audit-20260925`.

Q Sensitivity and Self-Consistency Comparison

These exploratory analyses reuse the existing records, with 10,000 paired item-bootstrap draws. Excluding Granite from the Section 6.1 task contrast leaves both remaining family contrasts positive after a two-comparison Bonferroni adjustment: +0.0648 [+0.0053, +0.1226] for Qwen3-4B and +0.1002 [+0.0328, +0.1690] for Phi-3.5-mini at 97.5% coverage. The point estimates equal the three-family rows because the contrast is computed per model.

The self-consistency signal is the fraction of an item's five E6b completions whose extracted numeric answer equals the greedy answer; a thin variant counts how many of the two sampled answers recorded in the original E6 streams equal the greedy answer. Each signal predicts greedy-answer correctness item by item and is paired against the verifier Yes/No score on the same greedy candidates.

Table Q1. Five-completion self-consistency agreement versus the verifier score on the same 700 GSM8K items per family.

Model	Verifier AUROC	SC-agreement AUROC	Difference [95% CI]
Qwen3-4B	0.8426	0.9094	+0.0668 [+0.0306, +0.1035]
Phi-3.5-mini	0.7494	0.9434	+0.1940 [+0.1578, +0.2316]
Granite-3.3-2B	0.3730	0.8985	+0.5255 [+0.4764, +0.5732]

Table Q2. Thin two-sample agreement from the original E6 streams versus the verifier score on the same greedy candidates, pooled over both windows.

Model	Verifier AUROC	SC-agreement AUROC	Difference [95% CI]
Qwen3-4B	0.8426	0.7971	-0.0455 [-0.0860, -0.0027]
Phi-3.5-mini	0.7521	0.8707	+0.1186 [+0.0770, +0.1622]
Granite-3.3-2B	0.3730	0.7987	+0.4257 [+0.3713, +0.4799]

Majority-vote accuracy exceeds greedy accuracy in all three families, at both the thin three-vote depth (0.4957, 0.7300 and 0.4871 versus 0.4329, 0.7000 and 0.4700) and the five-candidate depth (0.5986, 0.8014 and 0.5857): agreement both ranks correctness and improves the modal answer. The thin agreement signal trails the verifier score for Qwen3-4B, so the comparison depends on the number of samples.

Table Q3. Sixteen-sample agreement from the T4-3 resample (12 additional sampled completions per item, pool and sampling rule specified before generation) versus the verifier score on the same greedy candidates.

Model	Verifier AUROC	SC-agreement AUROC	Difference [95% CI]
Qwen3-4B	0.8426	0.9622	+0.1197 [+0.0870, +0.1527]
Phi-3.5-mini	0.7494	0.9715	+0.2221 [+0.1848, +0.2597]
Granite-3.3-2B	0.3730	0.9357	+0.5627 [+0.5159, +0.6085]

At sixteen samples the agreement signal dominates the verifier score in all three families with intervals excluding zero, the margin widens over the five-sample rows of Table Q1, and seventeen-candidate majority accuracy reaches 0.7386, 0.8729 and 0.6629.

R TruthfulQA Evaluation

We specify the protocol and question pool before evaluation and estimate policy contrasts from one run per model. The pool draws 684 four-option items from the TruthfulQA MC1 validation set (at least four options required), keeping the single true answer plus three seeded false answers with a fixed shuffle; gold positions are balanced at 170/177/168/169. Policies, readout and bootstrap match the MMLU protocol.

Table R1. TruthfulQA MC1 per-model AUROC, policy gaps and collision rates on 684 items.

Model	Chosen	Uniform-all	Δall [95% CI]	Δnon [95% CI]	Coll.	Coll.pos
Qwen3-4B	0.5656	0.7340	+0.1683 [+0.1170, +0.2208]	+0.0737 [+0.0065, +0.1396]	0.2368	0.6211
Phi-3.5-mini	0.5040	0.6767	+0.1726 [+0.1212, +0.2258]	+0.1009 [+0.0285, +0.1722]	0.2222	0.5901
Granite-3.3-2B	0.5644	0.6118	+0.0473 [-0.0080, +0.1024]	+0.0080 [-0.0570, +0.0738]	0.2456	0.4783

First-pass accuracies are 0.6316, 0.6155 and 0.4474. Granite-3.3-2B's chosen AUROC is 0.5644, compared with 0.6118 under uniform-all weighting; its policy-gap interval includes zero.