CAN CORRECTNESS READOUTS CHOOSE THE RIGHT ANSWER? A QUESTION-LEVEL AUDIT OF COMPACT KV-CACHE SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a correctness readout that ranks responses well in aggregate identify which answer to choose within a question? We study this question with KV24, a compact 24-feature summary of late-layer key–value cache statistics, paired with a lightweight correctness predictor; the language model remains frozen. We read these features before the dedicated final-answer field, although reasoning may already contain the answer. On held-out GSM8K, assigning every Qwen response its question's mean score removes all within-question ordering. The area under the receiver operating characteristic curve (AUC), computed over responses pooled across questions, changes from .9166 to .9104. Within-question ranking evidence is positive for Llama but inconclusive for Qwen. Hidden-state and likelihood controls match or outperform KV24 under several comparisons. Without target refitting, the joint Qwen–Llama ranking test passes narrowly on OpenBookQA but not on a synthetic cross-domain suite. Changing the source-validation selection criterion yields no consistent answer-choice benefit across models, and same-pool majority voting matches or exceeds the selected readouts. Together, these results separate boundary readability from transport and answer-choice utility.

1 INTRODUCTION

A language model can generate several candidate responses to the same question, each with its own reasoning trace and final answer. A correctness predictor assigns a scalar score to each response, with a higher score intended to indicate a greater chance of being correct. A system can then return the answer from the highest-scoring response. Internal model states are a promising source of such scores: prior work has recovered uncertainty and correctness signals from self-evaluation and hidden representations (Kadavath et al., 2022; Kossen et al., 2024; Zhang et al., 2025). The practical question is whether a score helps select a correct answer from the candidates available for a particular question.

To study this question, we use a fixed compact cache representation and ask how conclusions about correctness prediction change with the comparison unit, competing information, and downstream decision.

A common evaluation answers a broader question. It collects responses from many held-out questions and measures how often correct responses score above incorrect ones using the area under the receiver operating characteristic curve (ROC AUC). This *pooled* AUC includes comparisons between responses to different questions. A predictor can therefore perform well by assigning higher scores to responses from questions with high success rates and lower scores to responses from questions with low success rates. Such differences do not tell it which response to choose when one question has both correct and incorrect candidates. This is a question-level difficulty confound. For answer selection, the relevant distinctions are between responses to the same question, followed by the correctness of the answer actually chosen (Figure 1).

Concretely, we construct *KV24*: 24 fixed magnitude statistics of keys and values from late model layers, drawn from the key–value (KV) cache that a language model already maintains during decoding. We then train a lightweight logistic classifier on these features using response-correctness

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.

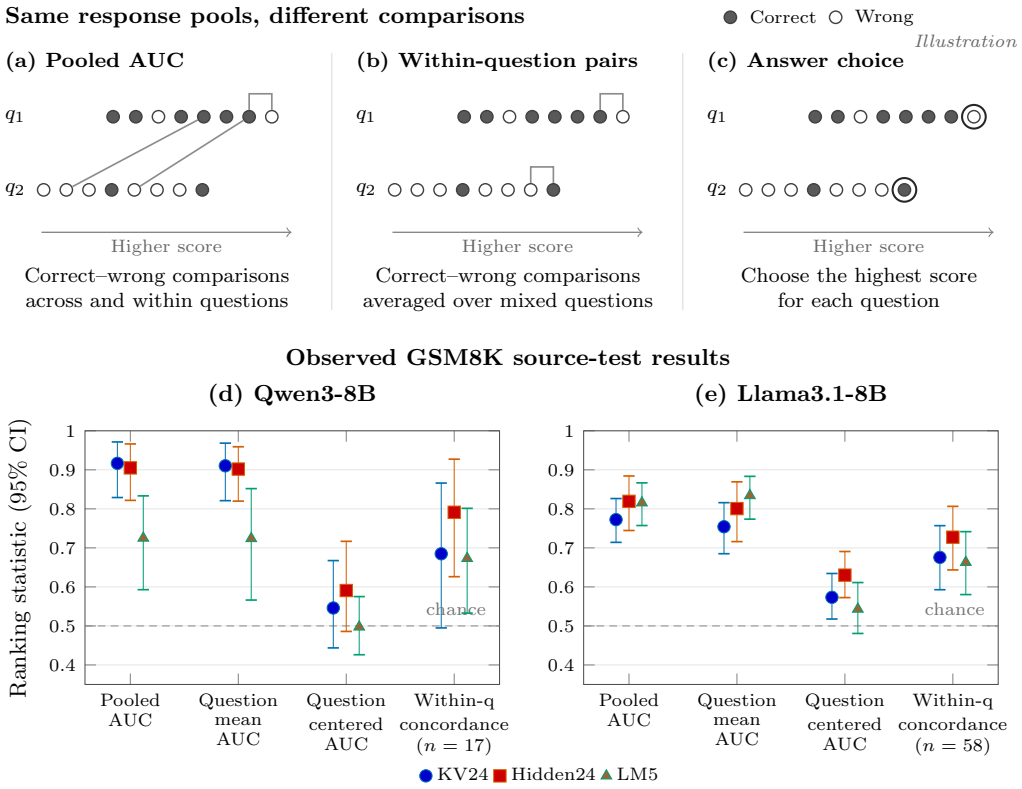

Figure 1: **High pooled AUC need not identify which answer to choose.** (a–c) The same response pools illustrate pooled and within-question comparisons and top-scoring answer selection. Filled/open circles denote correct/incorrect responses; rings mark selected answers. (d–e) Frozen GSM8K source-test sigmoid scores: question-mean replacement removes sibling ordering, whereas centering subtracts each question's mean. Bars are 95% percentile intervals from 5,000 question-clustered bootstrap draws; n counts mixed questions. Section 7 evaluates actual answer choice.

labels. The language model remains frozen. KV24 is the feature representation; the learned classifier is the *readout*, and its output is the correctness score. This distinction matters because the score is learned from labeled responses, rather than being the language model's own stated confidence. The small fixed feature set makes it possible to compare cache information with hidden-state and likelihood features under controlled fitting and evaluation conditions.

We extract all competing features after the final-answer marker and before the dedicated answer field. This *answer boundary* excludes subsequent answer tokens, but the preceding reasoning may already reveal the answer; it specifies what the readout observes, not when the model first determines an answer.

Prior work has identified question-level confounding in judge evaluation (Chen et al., 2026) and demonstrated answer-selection utility for hidden-state verifiers (Piotrowski et al., 2025). KV-CoE scores completed reasoning traces from value-cache trajectories (Xing et al., 2026), and OTV uses a learned cache-query token and LoRA verifier for answer selection and early termination (Zhuang et al., 2026). We connect these issues in a controlled audit of a compact cache readout: frozen-score decomposition, matched representation controls, source-frozen transfer, and direct answer-choice evaluation that accounts for source-side readout selection. Appendix Table 16 compares the observation and training interfaces.

We use *predictive advantage* to mean reliable improvement over tested hidden-state and likelihood controls under comparable conditions. This criterion separates signal readability from representation-specific value and differs from cross-model correctness prediction (Ashuach et al., 2026). We organize

the audit around four complementary questions: what is readable, whether KV24 offers a predictive advantage, whether its frozen score transfers, and whether it improves the intended decision. Together, these analyses separate readability from representation-specific and decision-level claims; Appendix Table 9 maps each claim to its supporting evidence.

2 Evaluation Questions and Estimands

Comparison unit and decision. Pooled ROC AUC is the probability that a randomly drawn correct response scores above a randomly drawn incorrect response (with half credit for ties), including cross-question pairs; unlike thresholded classification accuracy, it does not require a classification threshold. Within-question concordance compares pairs only within mixed questions (those with both outcomes), weighting questions equally. Answer choice evaluates each question's highest-scoring candidate, including questions with no eligible correct candidate. The comparison units and denominators differ (Figure 1a–c). Here *eligible* is the generic selection term; later tables use *valid* for boundary-valid attempts and *scorable* for attempts accepted by the dataset-specific grader.

Observed answer boundary. For question q_i, let B_{ij} contain the prompt and generated reasoning through the final-answer marker for response j, and let A_{ij} be the answer field that follows. Let $z(B_{ij})$ denote the feature vector extracted at that boundary—KV24 for the primary analysis, or the corresponding hidden or LM representation for a control—and let f be the fitted scalar correctness readout. Our score is $u_{ij} = f(z(B_{ij}))$. Two continuations from the same B receive the same score even if their answer fields differ. This readout cannot detect an error created only while realizing A; the reasoning in B may also already reveal the answer.

Frozen-score decomposition. For response j to question i, let u_{ij} be its saved, uncalibrated sigmoid correctness score, with n_i valid responses in its pool. We form

$$\bar{u}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} u_{ij}, \qquad u_{ij}^{\text{centered}} = u_{ij} - \bar{u}_i.$$

Assigning every response its question mean removes all sibling ordering; centering removes the question mean while retaining score differences among siblings. Both transformations keep the original fits fixed. They operate on sigmoid scores: converting to logits preserves rankings but changes means, centered comparisons, and the variance decomposition. Question means include the evaluated response and depend on the generated pool.

For mixed question i, let P_i and N_i index its eligible correct and incorrect responses. Its concordance is

$$C_i = \frac{1}{|P_i||N_i|}\sum_{j\in P_i}\sum_{k\in N_i}\left[\mathbf{1}\{u_{ij} > u_{ik}\} + \tfrac{1}{2}\mathbf{1}\{u_{ij} = u_{ik}\}\right].$$

The reported within-question statistic averages C_i equally over mixed questions. Centered AUC still compares residual scores across questions; concordance uses only sibling pairs. Questions lacking either outcome have no such pair and do not enter concordance, but remain in answer-choice accuracy.

Representations and transport. Cached keys and values are deterministic functions of the corresponding hidden activations and positions, so the associated data-processing argument concerns information availability rather than finite-sample learnability (Appendix A.3). We therefore test whether KV24's compression improves prediction over the controls, and whether its source-frozen score retains within-question ranking without target refitting.

Calibration and risk. Strictly increasing calibration preserves AUC and, under the same tie rule, fixed-coverage acceptance; accepted-risk guarantees additionally require a justified procedure and sampling assumptions (Section 7).

3 Evaluation Design

Compact KV readout. Primary runs capture the incremental cache at the last model-generated final-answer marker. KV24 summarizes magnitudes, discarding directions and head-specific structure;

conclusions concern this readout family. For a selected tensor T with M scalar entries, we compute

$$a(T) = \frac{1}{M}\sum_{r=1}^{M}|T_r|, \qquad e(T) = \sqrt{\frac{1}{M}\sum_{r=1}^{M}T_r^2}.$$

Three late layers, two suffix lengths (1 and 4 tokens), K/V, and these two reductions give $3\times2\times2\times2 = 24$ features. Source-training means and standard deviations yield $\tilde{z}$; ridge logistic regression predicts correctness as $\sigma(w^\top\tilde{z} + b)$. The base model stays frozen; correctness labels train the readout. Appendix Table 11 records the exact layer, tensor-axis, token-support, and fitting conventions. Appendix A details calibration and exact-token replay, which records zero KV drift on the audited GSM8K and MathQA captures.

What each control tests. The primary compact readouts share the ridge-logistic implementation. LM5 is a five-feature language-model confidence baseline measured at the same marker: next-token entropy; the mean, sum, and minimum token log-probability over the generated reasoning trace; and the number of scored trace tokens. KV24+LM5 tests incremental value beyond those features. Hidden24 is an equal-width hidden-state baseline: source-training labels select the 24 raw coordinates with the largest absolute standardized correct-minus-wrong mean difference, after which the same ridge-logistic readout is fit. Unlike KV24, Hidden24 uses labels to choose coordinates; it is equal-width but not matched in effective capacity. Hidden-Op12 tests the same aggregation operator on hidden states. Five-view analyses use view-specific compact LM/hidden summaries because token support changes across boundaries. KV-CoE is a label-free secondary comparator in its published raw direction (Xing et al., 2026).

Statistical units and primary tests. Questions—not individual responses—define data splits, out-of-fold folds, and bootstrap resamples. The OpenBookQA and cross-domain primary transport tests require the one-sided 95% question-bootstrap lower bound on within-question concordance to exceed .5 for both Qwen and Llama. Readouts and score directions are frozen before target outcomes are opened; controls and domain-specific rows are secondary. Appendix A.1 and Appendix A.4 give the full eligibility, denominator, and uncertainty rules.

Source-to-target sequence and evidence status. Qwen3-8B (Yang et al., 2025), Llama3.1-8B (Grattafiori et al., 2024), and Granite-3.3-8B each fit model-specific GSM8K readouts. Frozen Qwen and Llama readouts are evaluated without target fitting on OpenBookQA (Mihaylov et al., 2018) and the causal-science, instruction-following, and code suite. Granite is first evaluated on a separate model-specific target and then, descriptively, on the original cross-domain questions. This extension reuses the Qwen/Llama predictions rather than providing a new three-model confirmation. MathQA (Amini et al., 2019) provides additional calibration and sensitivity analyses. The source decomposition, representation comparisons, and answer-choice analyses are retrospective. The original OpenBookQA and cross-domain transport tests fix their readouts and decision rules before local target-outcome evaluation; Appendix A.1 records cohort roles and the evidence status of each analysis.

4 What Is Readable at the Answer Boundary?

The boundary is field-free, not answer-free. At the last marker, the dedicated answer field has not begun, but the reasoning may already reveal the answer. A literal-string audit finds this in nearly all GSM8K responses and a majority of MathQA calibration responses; literal absence is not evidence of semantic absence (Appendix B.1).

Readability rises most consistently between the prompt and the final marker (Table 1). Revealing the dedicated answer field produces a less uniform change: it improves KV discrimination on MathQA, but not clearly on GSM8K, while the LM and hidden summaries show no corresponding gain. Each view refits its readout, so the contrasts measure predictive availability, not a causal state transition; paired KV intervals appear in Appendix B.1.

Pooled readability is not response-specific ordering. Boundary readouts outperform measured prompt-only controls on the source test and the MathQA diagnostic, but the source late-KV increment

Table 1: AUC at three of the five audited boundaries (all five: Appendix Table 12). The question-disjoint out-of-fold (OOF) analysis uses 1,999 responses/250 GSM8K questions and 878 responses/878 MathQA-cal questions. Readouts share folds. LM-combined uses three common-support LM features; Hidden-summary uses seven label-free summaries, distinct from Hidden24/Hidden4096.

Cohort	View	LM-combined (3)	KV readout	Hidden-summary (7)
GSM8K	Prompt end	.5288	.5504	.4963
	Last marker	.7189	.7889	.6105
	Answer complete	.7077	.7936	.6120
MathQA-cal	Prompt end	.4927	.5421	.4667
	Last marker	.7858	.8220	.7339
	Answer complete	.7304	.8597	.7391

becomes inconclusive after conditioning on question means and available response controls (Appendix B.3). The limited prompt proxy therefore leaves response-specific KV attribution unresolved.

The frozen source-test scores show the distinction directly (Figure 1): replacing each KV score by its question mean preserves most pooled AUC in both models—for Qwen, .9166 to .9104—whereas subtracting that mean sharply reduces it. Centering leaves every within-question ordering unchanged. The drop in centered pooled AUC therefore reflects changed cross-question comparisons, rather than a loss of sibling-ranking ability. Sibling comparisons provide positive evidence for Llama but remain inconclusive for Qwen, which has only 17 mixed questions. Appendix Table 13 reports the point estimates.

5 Does KV24 Offer a Predictive Advantage?

Takeaway. Across equal-width, operator-matched, and higher-capacity comparisons, KV24 is readable but not consistently stronger than matched controls.

Table 2: Pooled AUC in the equal-width audit on boundary-valid common support. The source-test row uses train→calibration→test; full-source OOF uses five question-disjoint folds over 6,399 responses/800 questions; MathQA is source-refit transfer on target calibration. Random24 averages all 20 prespecified random-coordinate controls.

Cohort	KV24	Hidden24	PCA24	LM5	Random24 mean
GSM8K source test	.9166	.9050	.8543	.7251	.7822
GSM8K full-source OOF	.8202	.8554	.8422	.7161	.7409
MathQA calibration	.6735	.8532	.8187	.7930	.6870

Equal-width and operator-matched controls. KV24's small source-test lead over Hidden24 is not statistically resolved; full-source OOF evaluation reverses that ordering. Hidden-state and likelihood controls also outperform KV24 under MathQA development transfer (Table 2). Matching feature count does not ensure equal effective capacity.

The operator-matched audit uses the same layers, suffixes, statistics, split, and ridge pipeline for KV and hidden activations. The small pooled KV lead does not persist in the question-macro comparison or after adding common surface and LM controls; all paired difference intervals include zero (Table 3). Thus both equal-width and operator-matched controls leave KV24's comparative advantage unresolved.

LiLaVe-style boundary-matched XGBoost. A source-selected XGBoost readout at the same answer-field-free boundary has cohort-dependent ordering relative to KV24, with no resolved paired difference across four historical Qwen cohorts (Appendix Table 15; Piotrowski et al., 2025).

A separate MathQA check favors likelihood and full-hidden controls. Adding KV features to LM5 reduces discrimination, with the paired interval excluding zero (Appendix Table 34).

Table 3: Frozen fixed-test operator-matched hidden control from the 800-question source cohort. This protocol uses 32 responses per question and is complementary to, not row-identical with, the eight-response equal-width audit. Intervals resample questions.

Method	AUC [95% CI]	Macro concordance [95% CI]
KV24	.9168 [.8523,.9564]	.6870 [.5791,.7930]
Hidden-Op12	.8802 [.8235,.9234]	.7082 [.6118,.8044]
KV24+surface+LM	.9114 [.8399,.9603]	.7702 [.6852,.8449]
Hidden-Op12+surface+LM	.9216 [.8742,.9561]	.8224 [.7506,.8850]

6 Does the Frozen Readout Transport?

Takeaway. Frozen source readouts transport on OpenBookQA, while performance varies by target and matched controls often remain stronger.

6.1 OpenBookQA: narrow dual-model transport

Table 4: OpenBookQA equal-weight correct–wrong concordance within mixed questions. Brackets give frozen one-sided 95% lower bounds for KV24. KV-CoE-C denotes its primary published C score in the raw direction; it is a label-free secondary comparator. Invalid attempts count in all-attempt totals but are unscored.

Model	Valid/all	Mixed q	KV24 [95% LCB]	KV-CoE-C	Hidden24	LM5
Qwen3-8B	7,704/8,000	88	.5978 [.5433]	.4412	.6277	.5977
Llama3.1-8B	5,992/8,000	219	.5362 [.5041]	.4591	.5347	.6259

The frozen dual-model test passes because both prespecified one-sided lower bounds exceed chance (Table 4). Llama makes this a narrow success: its lower bound only just clears chance, its two-sided interval includes chance, a substantial share of attempts is invalid, and LM5 is stronger. This shows source-frozen transport without target fitting, but neither broad transport nor a KV-specific advantage.

Appendix D reports robustness and KV-CoE checks. Most retain the narrow result, whereas Llama's small complete-case subset does not.

6.2 Cross-domain suite: dual-model transport not established

The synthetic suite combines causal-science, instruction-following, and code tasks with deterministic gold answers or checkers. Task recipes and the generation budget were fixed on disjoint development questions.

Both Qwen and Llama meet the support requirements, but neither KV24 lower bound exceeds chance. The dual-model criterion is therefore unmet. On Qwen, KV24 also falls below Hidden24 and LM5 with both paired intervals below zero. Appendix Table 19 reports matched controls and domain rows.

Secondary domain point estimates place KV24 concordance above chance on science questions and below chance on code in both models, with instruction following near chance. These secondary point estimates suggest domain-dependent ordering, but do not isolate the cause of the differences (Appendix Table 19).

Sensitivity to domain weighting and common support. Equal weighting of the three domains gives KV24 intervals that include chance in both models. Restricting evaluation to questions mixed for both models puts Qwen's interval above chance while Llama's still includes chance. LM5 intervals stay above chance in both models under both sensitivities. These sensitivity comparisons show how domain weights and mixed-question populations affect the aggregate; neither alternative establishes positive KV24 transport in both models (Appendix Table 21).

Table 5: All models are evaluated on the same 500 cross-domain questions, with eight generation attempts per question and original strict grading. Scores are equal-weight mixed-question concordance; Δ is the paired difference between KV24+LM5 and LM5. Methods share response support within each model; mixed-question sets differ across models. For the prespecified Qwen/Llama rows, brackets in the KV24 column are one-sided 95% question-bootstrap lower confidence bounds and determine the primary pass/fail decision. Granite is a separately grouped descriptive extension and reports point estimates only. Qwen/Llama predictions are reused. Appendix Table 19 reports the original primary and domain results; Table 22 reports the matched-extension intervals.

Model	Scorable/all	Mixed q	KV24 [95% LCB]	Hidden24	LM5	KV24+LM5	Δ
Prespecified dual-model primary							
Qwen3-8B	3,990/4,000	321	.4721 [.4439]	.6711	.6213	.5708	−.0505
Llama3.1-8B	3,705/4,000	279	.5157 [.4896]	.5336	.5766	.5572	−.0194
Descriptive Granite extension (point estimates only)							
Granite-3.3-8B	2,240/4,000	238	.4856 (point)	.5903	.5014	.4881	−.0133

6.3 Granite on the original cross-domain questions

The Granite row in Table 5 extends its separate-target evaluation (Appendix D.4) to the original cross-domain questions. Here, KV24 is near chance, Hidden24 is stronger, and KV24+LM5 does not improve on LM5. Only 56.0% of attempts pass the original answer-format scorer; removing one outer backtick wrapper raises scorable coverage but leaves the overall KV increment unresolved. Code has too few mixed questions for a separate conclusion. These diagnostics expose format-sensitive coverage without isolating a model or interface effect (Appendix D.3).

7 Does the Readout Improve Answer Choice or Support Risk Control?

Takeaway. Better aggregate ranking does not automatically yield better answer selection; decision-matched evaluation changes which readout is preferred.

Metric choice changes the comparison. Concordance evaluates all correct–wrong pairs within a question; answer choice depends on which candidate ranks first. For illustration, the orderings C–W–W–C and W–C–C–W (highest score first; C/W denote correct/incorrect responses) both have concordance 0.5, but only the first selects a correct answer. Thus, even within-question concordance does not determine top-ranked answer correctness. Llama's cross-domain results illustrate the difference (Table 6). KV24 and Hidden24 both exceed LM5 in pooled AUC but select correct answers less often. LM5 has the highest concordance, while KV24+LM5 has the highest Best-of-8 accuracy. The ordering of these point estimates differs across all three metrics; Appendix Table 29 reports every target.

Table 6: Descriptive cross-domain comparison of all four frozen Llama readouts on 500 questions with eight attempts each. AUC uses eligible responses, concordance uses mixed questions, and Best-of-8 includes all questions.

Readout	Pooled AUC	Macro concordance	Best-of-8 accuracy
KV24	.6977	.5157	24.80%
Hidden24	.7034	.5336	25.40%
LM5	.6281	.5766	29.40%
KV24+LM5	.7532	.5572	30.20%

Answer choice under source-selected readouts. We compare two source-validation rules among the four frozen readouts: one maximizes pooled AUC, while the other maximizes Best-of-8 accuracy with AUC as a tie-breaker. This analysis changes how a readout is selected, not how it is trained.

Each selected readout returns the highest-scoring eligible response from a fixed budget of eight attempts. Invalid attempts remain in the budget, and all-invalid questions contribute zero accuracy. Majority voting uses the same candidate pool, and all questions remain in the accuracy denominator. The readouts score candidates individually at the answer boundary, whereas voting aggregates their completed answers. Thus, voting provides a decision-level baseline under the same generation budget, while boundary-matched feature controls assess representation-specific predictive value. Descriptive intervals reselect the method on source-validation resamples and independently resample evaluation questions, while keeping fits fixed. Table 7 reports the selected readout and target accuracy; Appendix E gives eligibility, tie-handling, and OpenBookQA subsampling details.

Table 7: Best-of-8 accuracy on 500 cross-domain questions per model under strict grading. Selection cells name the source-validation-selected readout and its target accuracy. Descriptive 95% intervals resample validation and test questions while keeping fits fixed. Llama's original selections coincide, but resampling the source-validation questions can make the two rules select different readouts.

Model	AUC-selected readout / accuracy	Best-of-8-selected readout / accuracy	Selection difference (Best-of-8−AUC), pp	Vote
Qwen3-8B	KV24+LM5 / 53.20%	Hidden24 / 59.80%	+6.60 [0.00,+10.80]	61.55%
Llama3.1-8B	Hidden24 / 25.40%	Hidden24 / 25.40%	0.00 [−7.60,+5.80]	32.32%
Granite-3.3-8B	Hidden24 / 33.00%	KV24+LM5 / 30.50%	−2.50 [−7.20,+2.10]	37.45%

With validation choices fixed, Qwen's cross-domain gain is 6.6 percentage points with a paired 95% interval of [2.4,11.0] (Appendix Table 28). Repeating the source-validation selection step changes the interval to [0.0,10.8] (Appendix Table 31), showing sensitivity to which readout the validation sample selects. Qwen's Best-of-8-selected readout gains only one validation question over the AUC-selected readout, and the two rules select the same readout in roughly half the resamples. On the cross-domain suite, the validation-and-test intervals comparing the Best-of-8-selected readout with voting include zero for Qwen and remain below zero for Llama and Granite (Appendix Table 31).

Selection baselines and candidate availability. For a question, let s_q be the original Best-of-8-selected policy's expected success, and let o_q indicate whether at least one eligible candidate is correct. Total error partitions as

$$1 - s_q = \underbrace{1 - o_q}_{\text{no eligible correct candidate}} + \underbrace{o_q - s_q}_{\text{eligible correct available but missed}} .$$

Under the current scoring and eligibility rules, the first term measures candidate availability and is unaffected by reranking; the second marks questions where a better selector could help. Both use all cross-domain questions as denominator, including all-invalid pools, and expected success retains uniform tie-breaking. This partitions observed errors without assigning their cause.

Table 8: Cross-domain answer-choice accuracy on 500 questions per model, with eight attempts each. Random selects uniformly among eligible responses; all-invalid pools score zero. Selected uses the source-validation Best-of-8 choices in Table 7. Eligible oracle is the fraction of questions with at least one eligible correct candidate. All entries are point estimates.

Model	Random	KV24	Selected	Vote	Eligible oracle
Qwen3-8B	51.54%	47.20%	59.80%	61.55%	83.80%
Llama3.1-8B	24.93%	24.80%	25.40%	32.32%	57.80%
Granite-3.3-8B	31.53%	27.90%	30.50%	37.45%	55.80%

Qwen's selected Hidden24 exceeds random choice in point accuracy but remains below voting; standalone KV24 is below random (Table 8). Llama's selected readout is close to random, while Granite's is lower. The eligible oracle averages o_q, so its gap from Selected gives selection error: 24.0, 32.4, and 25.3 percentage points for Qwen, Llama, and Granite. Appendix Table 32 reports the full error partition.

Selective filtering. Filtering can exploit between-question discrimination even when sibling ordering is weak. On a separate held-out Qwen GSM8K cohort, retaining the top 70% by KV24 score raises accepted accuracy from 90.84% to 97.14% (Appendix Figure 4). This fixed-pool result demonstrates filtering utility, but does not establish a transferable threshold or a prospective risk guarantee.

Risk-control diagnostic. Risk guarantees require a suitable calibration design, sampling assumptions, and control of multiple testing (Guo et al., 2017; Geifman & El-Yaniv, 2017; Bates et al., 2021; Angelopoulos et al., 2025).

The retrospective MathQA rule for KV24 accepts 1,294 of 1,847 eligible responses, with 117 errors (9.04%). Its conditional one-sided 95% upper risk bound is 10.46%, above the 10% target. This gap illustrates why an observed error rate below the target is not itself a risk certificate. A fixed-sequence learn-then-test analysis selects no KV24 score-quantile policy; a refit-pipeline sensitivity likewise extends above the target. Appendix Tables 36, 37, and 38 report the decisions, all-attempt accounting, and refit uncertainty.

8 Discussion and Limitations

Match the evaluation to the decision. Much of KV24's pooled discrimination reflects between-question variation, whereas sibling ordering depends on the model and target. Between-question separation can support pooled discrimination and selective filtering without identifying the best response within a question. Whether the frozen score preserves within-question ordering on a new target is a separate empirical question. Question-disjoint splits prevent leakage but do not turn pooled AUC into a within-question metric. Selection requires sibling ranking and final accuracy against same-pool voting, including source-side selection uncertainty (Table 7); filtering requires accepted risk at the target coverage. Each new target needs a frozen-score evaluation: source readability does not imply target transport. KV24's compact retained payload is a practical distinction: 96 bytes per response versus 655,360 bytes for the raw-hidden view, excluding extraction and generation costs (Appendix C.2).

General principle, empirical scope. The distinction among pooled discrimination, within-question ranking, and answer choice does not depend on the KV24 representation. Our comparative and transport findings, by contrast, concern the tested readouts, source fits, and candidate pools. These results illustrate a general evaluation requirement: measure answer-selection utility directly against same-pool baselines rather than infer it from pooled AUC.

What the audit establishes and what remains open. Math-domain additions are retrospective analyses; OpenBookQA uses a public key with a marginal Llama component; and matched Granite reuses opened questions, with only 99 mixed across all three models. The synthetic cross-domain suite is Qwen-developed, and answer formats and scorable coverage vary across models. Interface diagnostics change support and rankings without identifying the cause (Appendix D.5); literal audits miss paraphrases. Broader domain or architecture claims require matched new targets, while risk control requires a prospective calibration study.

9 Conclusion

We use KV24 as a compact test case to distinguish correctness discrimination, source-frozen transport, and decision utility. Much of its pooled signal survives removal of sibling ordering, and matched controls often perform as well or better. Transport succeeds narrowly on OpenBookQA but varies across targets; filtering improves accepted accuracy on a held-out Qwen cohort, while the selected readouts establish no advantage over same-pool voting.

The broader contribution is an audit organized around three requirements: match the statistical unit to the intended decision, compare against matched information sources, and evaluate transport with frozen source-to-target protocols. These requirements shift the focus from whether correctness is detectable to when a readable signal improves the intended decision.

REPRODUCIBILITY STATEMENT

The appendix and manifests document provenance, splits, fitting budgets, invalid-response accounting, and uncertainty procedures. The accompanying supplement includes standalone Python/NumPy replays. The cross-domain materials supply all 500 questions, deterministic gold/checkers, 200/200 science recomputation, a 30-question human audit, all 8,000 Qwen/Llama attempts, per-question outputs, and the evaluator; the documented command recomputes the original cross-domain estimates from saved predictions. The verifier selection materials supply normalized scores and code for metric-choice tables, reselection sensitivity, and error accounting. Included scripts reuse the source-test scores to reproduce Figure 1's estimates, intervals, and plot.

The anonymous number bundle includes the matched Granite, format-sensitivity, and source-validation metric-choice aggregate reports. These materials support the specified analyses and enumerated values, but do not regenerate every cohort and readout end to end (Appendix G).

ETHICS STATEMENT

This study evaluates model responses on public benchmarks and a synthetic question suite; it does not involve human subjects.

AI USE STATEMENT

Language models assisted research, code, editing, and surface drafting of cross-domain questions. Formal gold and grading are deterministic, with independent science recomputation and human audit; language models did not grade target responses. The authors verified the released materials and take responsibility for all claims.

REFERENCES

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. 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. 2357–2367, 2019.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Tomer Ashuach, Shai Gretz, Yoav Katz, Yonatan Belinkov, and Liat Ein-Dor. Masked by consensus: Disentangling privileged knowledge in llm correctness. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10577–10596, 2026.

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

Chenle Chen, Yangbo Wei, Chao Yao, Shaoqiang Lu, Junhong Qian, Chen Wu, and Lei He. Competence, not accuracy: A diagnostic for reference-free judge gates in skill optimization. *arXiv preprint arXiv:2608.18719*, 2026.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Shudong Liu, Hongwei Liu, Junnan Liu, Linchen Xiao, Songyang Gao, Chengqi Lyu, Yuzhe Gu, Wenwei Zhang, Derek F Wong, Songyang Zhang, et al. Compassverifier: A unified and robust verifier for llms evaluation and outcome reward. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33454–33482, 2025.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

Bartosz Piotrowski, Witold Drzewakowski, Konrad Staniszewski, and Piotr Miłoś. Lightweight latent verifiers for efficient meta-generation strategies. *arXiv preprint arXiv:2504.16760*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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. *arXiv preprint arXiv:2203.11171*, 2022.

Zeyu Xing, Xing Li, Hui-Ling Zhen, Mingxuan Yuan, and Sinno Jialin Pan. Beyond speedup–utilizing kv cache for sampling and reasoning. *arXiv preprint arXiv:2601.20326*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*, 2025.

Zhan Zhuang, Xiequn Wang, Zebin Chen, Feiyang Ye, Ying Wei, Kede Ma, and Yu Zhang. One-token verification for reasoning correctness estimation. *arXiv preprint arXiv:2603.01025*, 2026.

A Experimental Protocol

This appendix follows the main evaluation questions. Table 9 locates the supporting evidence; this section collects the shared methods.

Table 9: Guide to the supporting methods and results.

Question	Supporting detail	Location
What is readable at the answer boundary?	Boundary visibility and question–response decomposition	Appendix B
Does KV24 offer a predictive advantage?	Equal-width, operator-matched, and raw-hidden controls	Appendix C
Does the frozen readout transport?	Target protocols, domain support, and format sensitivity	Appendix D
Does the readout improve answer choice or support risk control?	Answer selection; selective filtering and risk	Appendices E, F

A.1 Data, generation, and correctness labels

Each model's main GSM8K source uses the same 800 questions, split into 480 training, 160 validation, and 160 test questions, with eight responses per question. Training fits transforms and readouts, validation calibrates and selects readouts, and test remains held out; readouts are then frozen. Splits, out-of-fold folds, and bootstrap resamples operate at question level, retaining sibling responses and paired method scores. Table 10 separates cohorts with different response budgets or fitting roles.

Qwen decoding uses its chat template with thinking disabled, temperature .9, top-p .95, and natural stopping; GSM-style generation is capped at 1,024 new tokens. Replications use model-native chat templates. Exact generation parameters, question identities, split assignments, and response seeds are retained in the accompanying manifests. Capped, missing-boundary, and invalid-format attempts remain in attempted denominators. Ranking uses eligible responses; selection includes all questions and counts all-invalid pools as failures. A format rejection does not itself establish a wrong content answer.

GSM-style answers use normalized numerical comparison, OpenBookQA uses option letters, and the cross-domain suite uses deterministic option/JSON checkers. Labels do not resolve score or voting ties. External questions are screened against project data using fixed exact and near-duplicate rules. For MathQA, one previously exposed official-test row is removed; a label-free partition allocates 900 calibration, 1,900 test, and 184 unused reserve questions. Historical teacher-forced and online-cache cohorts are evaluated separately.

Evidence status is summarized here. The matched source and representation comparisons, validation-selection analyses, and reanalyses of previously inspected targets were specified after outcome inspection and are treated as retrospective. OpenBookQA and the original cross-domain and separate Granite primaries fix readouts and analysis rules before joining target labels; their registrations are internal to the project. OpenBookQA's key is publicly available. The synthetic cross-domain questions have no normalized-text overlap with development, but use Qwen-developed task recipes. The same-question Granite extension is likewise descriptive.

A.2 Boundary capture and feature schema

The measured context ends at the generated final-answer marker, before the dedicated answer value and EOS. Reasoning can already contain the answer. Online runs save the actual incremental cache; exact-token incremental replay reproduces the saved KV24 values. BF16 full-forward execution can give numerically different states on the same token sequence, so teacher-forced features are not substituted for online captures.

Table 10: Cohorts used in the reported comparisons. Responses/q is the attempted budget per model. Shared questions do not make different response pools or fitting protocols interchangeable. Boundary-view and interface subsets are specified in their respective analyses.

Model	Cohort	Questions	Responses / question	Role
Main source and target pools				
Qwen / Llama / Granite	GSM8K source	800	8	Model-specific source fit; 480/160/160 split
Qwen / Llama	OpenBookQA	500	16	Fixed readouts/endpoints; public key; no target fit
Qwen / Llama	Cross-domain	500	8	Outcome-isolated dual-model primary; no target fit
Granite	Cross-domain	500	8	Descriptive extension on the same questions; no target fit
Supporting comparisons and sensitivities				
Qwen	GSM8K historical source	800	32	Hidden/operator controls on separate response pools
Qwen	External GSM8K	500	8×2	Held-out filtering and raw-hidden comparisons
Qwen	GSM1K	800	8×2	Raw-hidden and answer-frequency comparisons
Qwen	GSM-Hard development	500	$\leq 8 \times 2$	Target scalar calibration and threshold fitting
Qwen	GSM-Hard reserve	187	$8 + 1$	Held-out evaluation; later comparisons are retrospective
Qwen	MathQA calibration	900	1	Target scalar/threshold fitting and risk diagnostics
Qwen / Llama	MathQA test	1,900	1	Qwen historical frozen-rule test; Llama source-only transfer
Granite	Separate target	300	8	Separate outcome-isolated primary; no target fit

Corrected scoring uses only the 24 base features. A redundant aggregate in historical artifacts is excluded. The supplementary reports retain the schema correction and replay checks; neither changes the boundary definition.

A.3 Features, readouts, and calibration

Table 11 fixes the implementation choices needed to reconstruct every named feature family. Transformer-block indices and hidden coordinates are zero-based throughout.

Table 11: Feature and readout specification for the reported analyses. "Source train" and "source calibration" refer to question-disjoint GSM8K splits; target labels do not alter a source-frozen representation or readout.

Object	Exact construction and fitting rule
KV24	Use the model-native incremental `past_key_values` at the last generated marker: Qwen blocks $\{31, 32, 33\}$, Llama $\{27, 28, 29\}$, and Granite $\{35, 36, 37\}$. The cached keys are post-rotary-position transform in these model implementations. Each cached tensor has axes batch $\times$ KV head $\times$ sequence $\times$ head dimension; absolute mean and RMS reduce over batch, every KV head, the selected suffix positions, and head dimension. Suffixes $S \in \{1, 4\}$ are the last one or four generated response tokens. Both end at and include the token whose decoding completes the final marker; $S = 4$ need not equal the marker's full tokenization. The ordered product layer $\times$ suffix $\times\{K, V\} \times \{\text{absolute mean}, \text{RMS}\}$ gives 24 values.
Hidden24	Qwen first averages 4096-D hidden states over blocks $\{20, 28, 32, 34, 35\}$ and the final 16 generated response tokens ending at the marker. Llama and Granite retain a separate final-16-token mean for every block; source-train selection chooses block 17 for Llama and block 20 for Granite. From the resulting 4096-D vector, source-train labels select the 24 coordinates with largest absolute standardized correct-minus-wrong mean difference, then the common standardized ridge-logistic readout is fit. In the five-fold question-OOF Qwen analysis, coordinate selection, standardization, and the readout are repeated inside every training fold; the frozen transfer model is refit once on the designated full source-fit cohort. The Llama and Granite source replications perform the train-only layer and coordinate selection once, not per target.
Hidden4096	This full-coordinate control uses the same boundary-hidden construction as Hidden24 but retains all 4,096 coordinates. Qwen averages blocks $\{20, 28, 32, 34, 35\}$ over the final 16 generated response tokens ending at the marker. Llama and Granite use the full vector from the source-train-selected Hidden24 block (17 and 20, respectively), again averaged over the final 16 tokens, or all available tokens without padding. Coordinate standardization and the ridge-logistic readout (penalty 1.0) use source-training rows only; the reported Llama and Granite models are fixed-model conditional controls and are not reselected on target data.

Continued on next page

Table 11 (continued)

Object	Exact construction and fitting rule
PCA24 / Random24	These Qwen equal-width controls start from the same 4096-D Qwen hidden mean used by Hidden24. PCA24 standardizes coordinates on source-training rows and uses randomized SVD with 24 components, eight oversamples, one power iteration, deterministic sign orientation, and seed 2600064442 for the last-marker view; projected scores are *not* whitened. PCA fitting is repeated inside each OOF training fold. Random24 samples 24 distinct raw coordinates without replacement for each fixed seed: 1729, 3253, 4903, 6571, 8231, 9901, 11579, 13249, 14923, 16603, 18287, 19973, 21661, 23357, 25057, 26759, 28463, 30169, 31873, and 33581. The reported control is the mean over all 20 seeds; no seed is selected, and the coordinate subsets remain fixed across OOF folds.
LM5	For every generated continuation token through and including the complete marker, compute its log-probability under the model's raw full-vocabulary distribution before temperature scaling or top-p truncation. Prompt tokens, the later answer field, and EOS are excluded. LM5 contains the mean, sum, and minimum of these token log-probabilities, their count, and the full-vocabulary entropy of the raw next-token distribution immediately after the marker. The same standardized ridge-logistic implementation used for KV24 maps these five features to a score.
KV-CoE	Exact-token incremental replay builds $e_t = L^{-1}\sum_l \mathrm{concat}_h V_{lht}$ from every value-cache layer over the full prompt-plus-response context through the marker. Let $m_t = \|e_t - e_{t-1}\|/\|e_T - e_1\|$ and $a_t = \arccos(\cos(e_t, e_{t-1}))\,/\,\arccos(\cos(e_T, e_1))$. The reported primary score is $C = \sqrt{(\overline{m_t \cos a_t})^2 + (\overline{m_t \sin a_t})^2}$; secondary $R = \bar{m} - \bar{a}$. Layer means accumulate in FP32, cosines are clipped to $[-1, 1]$, denominators must exceed 10^{-12}, and degenerate trajectories are unscored. This independently implements Eqs. (6)–(9) of Xing et al. (2026) and the endpoint normalization in released `score.py` commit `41d469967146`; it has no fitted readout and is evaluated in the published raw direction.
LiLaVe-style XGBoost	The fair row in Table 15 applies a 4096-D statewise XGBoost classifier to Qwen blocks $\{20, 28, 32, 34, 35\}$ at the final 16 response positions, then averages the 80 state probabilities. The objective is `binary:logistic` with histogram trees, learning rate 0.1, and seed 370; source calibration chooses between depth 5/100 rounds and depth 10/30 rounds by AUC, then Brier score, model size, and identifier. Thus the faithful table has a two-candidate selection budget and a fixed location. A separate deployment sensitivity searches both recipes across five registered location masks (10 candidates) and is not substituted for that row. This historical comparison uses the Qwen S32 source pool and completion-valid literal-prefix rows from its associated external pools, rather than the primary online-native eight-response pool; it is therefore a capacity/observability control, not a matched rerun of the main experiment.

Information availability and learnability. Let $\mathcal{S}$ contain every layer–token pair used by a representation, $H_{\mathcal{S}}$ the corresponding joint hidden activations, and $P_{\mathcal{S}}$ the corresponding positions or fixed positional transforms. Under a fixed Transformer (Vaswani et al., 2017), the joint cached keys and values $(K_{\mathcal{S}}, V_{\mathcal{S}})$ are deterministic functions of $(H_{\mathcal{S}}, P_{\mathcal{S}})$. Therefore, for any summary ϕ,

$$I(Y; \phi(K_{\mathcal{S}}, V_{\mathcal{S}})) \le I(Y; H_{\mathcal{S}}, P_{\mathcal{S}})\,.$$

This inequality concerns information availability, not finite-sample learnability. Compact KV summaries may still offer useful compression or inductive bias. Neither the inequality nor an unresolved difference under the tested controls establishes equivalence among all KV and hidden readouts.

Readout fitting and calibration. For ridge-logistic probes, a scaler is fit on training questions only and the ridge penalty is 1.0 unless a registered protocol states otherwise. Ranking uses uncalibrated source scores (logits or their sigmoid probabilities). A positive-slope Platt map, fit only on calibration questions, estimates $P(\text{correct})$; predicted risk is one minus this estimate. Consequently AUC, area under the risk–coverage curve (AURC), rank-matched coverage, and concordance use the source ordering, while Brier score, negative log-likelihood (NLL), expected calibration error (ECE), and probability thresholds use the calibrated probability. Figure 1 retains saved sigmoid scores for means and centering (Appendix B.2). Exact-coverage curves accept the top $\lceil cn \rceil$ responses by the frozen ranking. Threshold-transfer experiments fix cutoffs on the designated calibration split. Separately labeled outcome-free quantile diagnostics instead use target score distributions without correctness labels. Under GSM-Hard shift, only a scalar monotone target map and its development threshold may be fit; the source feature scaler and source probe remain frozen. The raw-hidden comparator fits its source Platt map with positive-slope bounded L-BFGS, checking finiteness and convergence.

Five-view diagnostics reconstruct these operators at each named boundary; target outcomes do not alter a source-frozen transform, score direction, or readout.

A.4 Estimands and clustered uncertainty

We bootstrap questions, retaining all sibling responses, labels, and paired method scores in each draw. Unless a protocol declares a larger count, we use 2,000 replicates and percentile 95% intervals. Pooled AUC estimates the probability that a randomly drawn wrong response has higher risk than a randomly drawn correct response. Question-macro concordance first computes within-question correct–wrong pair orderings, then weights questions equally. The historical within-question top-1 diagnostic conditions on pools containing both outcomes. Best-of-8 in Section 7 instead includes all questions.

Repeated cross-fitting adds a second dependence axis. Its primary number is the mean of five independently evaluated question-OOF repeats; uncertainty resamples realized repeat and question clusters without retraining. It is therefore conditional on those five fitted repeats, not an interval over all possible training sets. Selective-prediction intervals either hold the frozen calibration threshold fixed or, when explicitly labeled, independently resample calibration and test questions and refit only the threshold. Neither procedure refits the representation probe.

B Boundary Readability and Score Decomposition

These analyses ask which information is available at the measured boundary and how pooled discrimination relates to within-question ordering.

B.1 Answer visibility and five-view replay

The reasoning often reveals the answer before the dedicated field begins. Literal matching finds the realized value in 96.45% of audited GSM8K captures and the option value or reference in 61.39% of MathQA calibration captures. These literal matches establish visibility; literal absence does not establish semantic absence or provide a deployable filter.

All three readouts improve from prompt end to the last marker on both cohorts (Table 12); paired 95% intervals exclude zero. Generating the dedicated answer field adds no resolved gain on GSM8K. On MathQA calibration, it improves KV discrimination but weakens the LM summary, with no resolved hidden-summary change. Thus most of the consistent gain appears before the answer field. These paired intervals use 5,000 question resamples; refitting each view measures predictive availability, not a causal effect of exposing the answer. For the last-marker-to-answer-complete KV contrast, the paired AUC changes are .0048 [$-.0616$,.0694] on GSM8K and .0377 [.0091,.0695] on MathQA calibration. The answer-visibility report in the supplementary materials provides the complete paired intervals for every readout and contrast.

Table 12: Complete five-view question-disjoint OOF AUC on GSM8K (1,999 responses/250 questions) and MathQA-cal (878/878). Models are refit per view on shared folds; intervals condition on fixed OOF predictions. The compact three- and seven-feature controls are not LM5, Hidden24, or Hidden4096.

Cohort	View	LM-combined (3)	KV readout	Hidden-summary (7)
GSM8K	Prompt end	.5288	.5504	.4963
	First marker end	.7189	.7885	.6103
	Last marker end	.7189	.7889	.6105
	Answer complete	.7077	.7936	.6120
	Post-answer	.7027	.7757	.6196
MathQA-cal	Prompt end	.4927	.5421	.4667
	First marker end	.7892	.8112	.7323
	Last marker end	.7858	.8220	.7339
	Answer complete	.7304	.8597	.7391
	Post-answer	.7340	.8561	.7383

B.2 Question-versus-response decomposition

The decomposition follows the probability-scale definitions in Section 2. The same 160 source-test questions are resampled 5,000 times for both models and all methods, with scores and fits frozen. Table 13 gives point estimates; Figure 1 gives intervals.

Table 13: Frozen-score statistical-unit decomposition. Q-mean assigns each valid response its question's mean score; Q-centered subtracts that mean; Within-q averages correct–wrong concordance over 17 mixed Qwen and 58 mixed Llama questions. Between-q var. is the fraction of score variance attributable to question means.

Model	Method	Pooled	Q-mean	Q-centered	Within-q	Between-q var.
Qwen	KV24	.9166	.9104	.5460	.6849	.7508
	Hidden24	.9050	.9019	.5907	.7913	.8347
	LM5	.7251	.7240	.4974	.6730	.5427
Llama	KV24	.7724	.7542	.5733	.6756	.5959
	Hidden24	.8193	.8006	.6301	.7274	.5947
	LM5	.8154	.8341	.5431	.6629	.5091

Question means retain much of pooled KV24 discrimination in both models (Table 13). Centered and within-question intervals exclude chance for Llama but remain unresolved for Qwen, whose mixed support is smaller (Figure 1). Cross-model correlations in question accuracy and mean KV scores indicate shared question structure without making the readout a model-independent question index.

B.3 Prompt-conditioned controls

Beyond prompt+response controls (Table 14), the pooled gain is .085 [.023,.147]; the question-centered and question-macro intervals both include zero. MathQA calibration likewise provides no resolved increment beyond the observed control family. The historically locked MathQA comparison still favors KV over its limited prompt-surface proxy, which cannot rule out unmeasured semantic question structure or isolate a causal response signal.

Marker choice. First and last markers coincide for single-marker responses. Replacing the last marker with the first changes overall AUC little in the audited cohorts; the repeated-marker strata are small. The supplementary first-marker report gives the complete sensitivity.

Table 14: Prompt-conditioned sensitivity on 5,120 responses from 160 source test questions. Q-centered and Q-macro measures require repeated responses and cannot be computed for one-response-per-question MathQA.

Source fixed-test model	Pooled AUC	Q-centered AUC	Q-macro AUC	Between-Q fraction
Prompt controls only	.513	.500	.500	1.000
Prompt context only	.660	.500	.500	1.000
Prompt+response controls	.767	.642	.714	.758
Prompt+response+late KV	.852	.683	.808	.783
Paired response controls	.730	.645	.728	.767
Paired response+late KV	.788	.642	.736	.748

C Representation and Readout Comparisons

The controls vary feature width, aggregation operators, or readout capacity. Their paired comparisons use the same eligible rows within each analysis; historical response pools remain separate from the main eight-response source.

C.1 Equal-width controls and source OOF stability

On the same 160-question source test, the label-free KV-CoE-C score reaches .6737 [.4974,.8290] AUC and KV-CoE-R reaches .6565 [.4828,.8204]. Both use their published direction; neither interval excludes chance. The learned controls below separately address supervised readout comparisons.

Coordinate selection, PCA, standardization, and readouts use source-training rows only; scalar calibration uses a separate split. On the fixed test, KV24's differences from Hidden24 and PCA24 are unresolved after Holm correction, although it exceeds the mean of all 20 Random24 controls. The Hidden24-minus-KV24 interval is $[-.0518, .0265]$.

Full-source question-disjoint OOF places Hidden24 above KV24, while hidden and likelihood controls are also stronger under MathQA transfer (Table 2). Replacing the original fits with OOF fits on the same source-test rows also reverses that ordering. Most of KV24's gap between the original test and full-source OOF is descriptively associated with cohort composition; the same-cohort fit-change interval includes zero. This sensitivity limits interpretation of the favorable original split.

All four operator-matched KV-minus-hidden intervals also include zero, across pooled and question-macro comparisons with or without surface/LM controls (Table 3).

C.2 LiLaVe-style XGBoost comparison

Table 15: LiLaVe-style XGBoost comparison on common completion-valid, literal-prefix rows. Boundary XGB excludes answer-value and EOS tokens; full-response XGB observes answer-value tokens. The paired AUC difference is boundary XGB minus KV24 and clusters by question.

Cohort	Rows	Boundary XGB	KV24	Δ [95% CI]	Full-response XGB
Source test	5,120	.931	.917	+.014 [−.017,.050]	.941
External GSM8K	7,999	.810	.822	−.011 [−.050,.025]	.817
GSM1K	12,798	.834	.804	+.030 [−.001,.063]	.857
GSM-Hard	1,494	.687	.718	−.031 [−.082,.021]	.726

The boundary-matched XGBoost comparison gives no consistent KV24 advantage (Table 15). Giving the XGBoost readout answer-value tokens improves its discrimination on GSM1K and GSM-Hard. In the opened GSM-Hard reserve, identical target calibration likewise leaves the boundary-XGBoost versus KV24 differences unresolved. These observability controls are separate from the earlier prospective calibration evaluation.

KV24 has a much smaller retained payload: 96 bytes per response versus 655,360 for this raw-hidden view. That comparison excludes extraction and generation. The released timing tables measure post-hoc teacher-forced extraction and frozen scoring, not online generation-hook overhead.

Table 16: Closest internal-state verifiers differ in representation, adaptation, and readout timing. We audit a fixed KV feature family across evaluation units and source-frozen targets.

Method	**Representation**	**Adaptation**	**Readout timing**	**Evidence target**
LiLaVe (Piotrowski et al., 2025)	Selected hidden states	Frozen base; learned readout	Completed response	Best-of-N; weighted voting; cross-dataset transfer
Hidden verification (Zhang et al., 2025)	Intermediate hidden states	Learned probe	During reasoning	Future correctness; early exit
OTV (Zhuang et al., 2026)	KV queried by learned token	LoRA verifier role	Anytime query	Best-of-N; weighted voting; early termination
KV-CoE (Xing et al., 2026)	Token-centric value-cache trajectory	Frozen formula; no fitted readout	Completed trace	Self-evaluation
Ours	24 aggregate K/V statistics	Frozen base; ridge probe	Generated final-answer marker	Pooled versus same-question ranking; answer choice; transport

C.3 Additional information budgets

Sampling-based confidence (Wang et al., 2022; Kuhn et al., 2023) uses multiple responses, whereas KV24 scores one response at the boundary. The following comparison makes that budget difference explicit.

Table 17: Label-free numeric answer-group frequency versus corrected 24-feature KV. The frequency baseline uses eight responses; KV scores one. Intervals cluster by question.

Cohort	Answer frequency (8 samples)	KV (1 sample)	Δ AUC [95% CI]	Δ macro [95% CI]
External GSM8K	.852	.822	+.031 [−.036,.098]	+.104 [−.006,.206]
GSM1K	.904	.804	+.101 [.063,.140]	+.082 [.003,.161]
GSM-Hard	.818	.718	+.100 [.048,.159]	+.125 [.006,.243]

Answer frequency improves ranking on GSM1K and GSM-Hard, but uses eight responses rather than one (Table 17). Its higher point AUC on external GSM8K accompanies worse AURC, so it does not uniformly improve selection. Pool-level consensus and inverse entropy are constant within an episode and cannot rank its responses.

The supplementary materials also retain the gold-aware CompassVerifier comparison (Liu et al., 2025), hidden-readout label-efficiency curve, and earlier statistical-unit diagnostics. Gold-aware verification has a different information budget; the label-efficiency curve characterizes the hidden comparator rather than KV24. Neither comparison determines the matched KV24 conclusions above.

D Source-Frozen Transport and Sensitivity

The target evaluations follow the main text: OpenBookQA, the cross-domain suite, and the matched Granite extension. Supporting math-domain and separate-target evaluations follow, with their distinct evidence status.

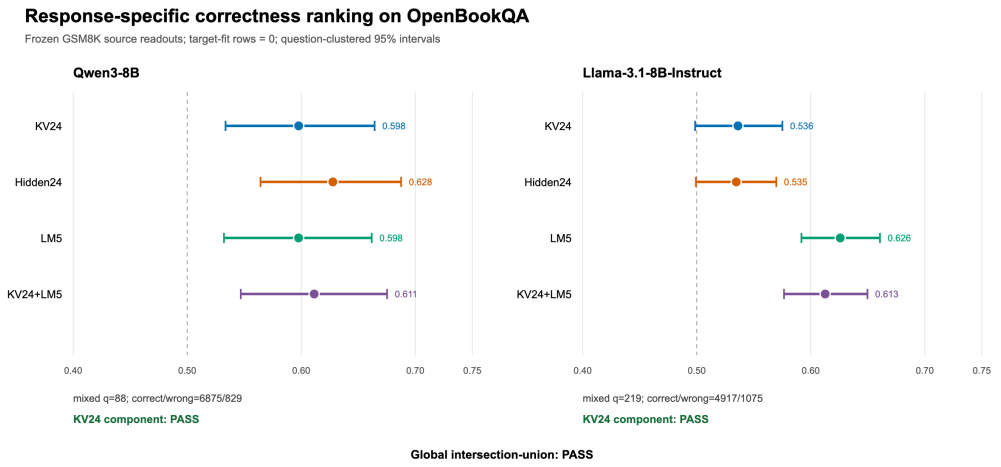

Figure 2: OpenBookQA response-specific evaluation. Points are equal-weight within-question correct–wrong concordance; bars are two-sided 95% question-bootstrap intervals. The dashed line marks chance. The confirmatory decision uses the prespecified one-sided lower bounds for KV24, not whether each displayed two-sided interval excludes .5. All readouts are frozen from GSM8K.

D.1 OpenBookQA

The protocol fixes 500 official test questions, 16 responses per question and model, four source-frozen readouts, and no target fitting. Both score bundles were sealed before joining the public key. Invalid attempts remain in accounting without imputed scores or labels. Qwen retains 7,704/8,000 valid attempts and 88 mixed questions; Llama retains 5,992/8,000 and 219. The primary requires each model's one-sided 95% question-bootstrap lower bound for KV24 within-question concordance to exceed .5. Both components pass, giving an intersection–union test $p = .02859$.

Llama is a boundary result: its two-sided interval includes chance, while the prespecified one-sided lower bound is .5041 (Figure 2). Neither model establishes an advantage over Hidden24, and Llama LM5 is stronger. Independent statistical reimplementation exactly reproduces all bootstrap draws and the primary decision.

A subsequent comparison tests KV-CoE at the same boundary. Both KV-CoE variants have lower bounds below chance, and KV-CoE-C is below KV24 in both models. This uses their published score directions without source or target fitting. The comparator uses incremental decoding because single-pass full-forward replay is not score-equivalent.

The sensitivity audit in Table 18 conditions on the observed mixed-question set. Label permutation is performed within each question while preserving its valid-response and correct/wrong counts. BCa acceleration is obtained by deleting one mixed question at a time; bootstrap-t uses the resampled question-level standard error.

Table 18: OpenBookQA robustness audit with 50,000 draws. The first three columns and valid-weighted column are one-sided 95% lower bounds. The complete-case lower bounds are .5946 and .3646, respectively.

Model	Conditional	BCa	Boot-t	Perm. p	Valid-wtd. LCB	Complete q / point
Qwen3-8B	.5423	.5416	.5403	$< .00002$	.5494	59 / .6624
Llama3.1-8B	.5043	.5038	.5039	.01588	.5014	16 / .4771

Alternative bootstrap intervals, within-question permutation, and valid-response weighting retain the narrow positive result (Table 18). Complete-case restriction does not: Llama is left with only 16 fully valid question pools. These sensitivities show dependence on response validity.

D.2 CROSS-DOMAIN SUITE

The cross-domain suite uses 500 new questions, eight responses per model, and four frozen GSM8K readouts. Both models meet the prespecified support minimum of 75 mixed questions and 150 valid responses of each correctness class. Yet neither KV24 component passes the lower-bound-above-chance criterion (Table 19). Hidden and likelihood scores retain signal overall, while KV24 reverses direction on code.

Table 19: Cross-domain suite equal-weight within-question concordance. Rows labeled *primary* give the frozen all-domain KV24 values and one-sided lower bounds; domain rows and comparator columns are secondary. Qwen retains 3,990/4,000 scorable attempts and Llama 3,705/4,000.

Model	Domain	Mixed q	KV24 [95% lower bound]	Hidden24	LM5
Qwen3-8B	*All (primary)*	321	.4721 [.4439]	.6711	.6213
	Science	97	.6514	.6654	.6184
	Instruction	93	.5232	.6544	.6629
	Code	131	.3030	.6871	.5940
Llama3.1-8B	*All (primary)*	279	.5157 [.4896]	.5336	.5766
	Science	176	.5480	.5959	.5944
	Instruction	53	.5128	.5472	.5781
	Code	50	.4049	.2997	.5128

Paired comparisons reinforce the contrast with controls: KV24 is below Hidden24 and LM5 in Qwen, and below LM5 in Llama. Adding KV24 also reduces LM5 concordance in both models. These secondary comparisons use the frozen score directions, including the code-domain reversal.

Table 20: Cross-domain suite domain support. Accuracy conditions on scorable responses; invalid attempts remain in the displayed attempted denominators. The Qwen-tuned recipes place Llama instruction and code well below the intended capability boundary.

Model	Domain	Valid/all	Accuracy	Mixed q	Invalid
Qwen	Science	1,598/1,600	.6859	97	2
	Instruction	1,196/1,200	.2458	93	4
	Code	1,196/1,200	.5577	131	4
Llama	Science	1,529/1,600	.4971	176	71
	Instruction	1,091/1,200	.1072	53	109
	Code	1,085/1,200	.0599	50	115

Table 21: Estimand sensitivities. Equal-domain rows average the three domain concordances after resampling mixed questions within domain. Joint-mixed rows use the 169 questions mixed for both models and resample common identities. Intervals are two-sided 5,000-draw percentile question bootstraps.

Model	Alternative weighting	Support	KV24 [95% CI]	Hidden24	LM5
Qwen	Equal domain	97/93/131	.4925 [.4619,.5245]	.6690	.6251
	Joint mixed	169	.5508 [.5014,.5963]	.6530	.6384
Llama	Equal domain	176/53/50	.4886 [.4501,.5246]	.4810	.5617
	Joint mixed	169	.4962 [.4550,.5389]	.4846	.5413

Llama Hidden24 is less stable under the weighting and common-support sensitivities in Table 21. The Qwen-developed recipes also place Llama instruction and code below the intended capability boundary (Table 20). These differences limit a model-only interpretation.

Before generation, the formal question bundle passes checker self-tests, independent recomputation of its science mechanisms, unique-option checks, and a sealed human-sample audit. Question-only

GPU inputs exclude gold and checkers; both prediction bundles are verified before the first gold join. No question exclusion, sign reversal, or scorer refit is selected after opening outcomes.

D.3 GRANITE ON THE SAME CROSS-DOMAIN QUESTIONS

This extension adds Granite responses to the original 500 cross-domain questions, reuses the Qwen/Llama predictions, and keeps Granite's GSM8K-source readouts frozen. Task wording, answer contracts, and the eight-response budget are unchanged; generation uses a 1,024-token cap, temperature .9, top-p .95, and model-native chat templates. The designated contrast is KV24+LM5 minus LM5. Because the questions were already opened, these results are descriptive and remain separate from the earlier frozen primaries.

Estimand and support. Methods share strictly scorable responses within each model. Concordance averages mixed questions, whose identities can differ across models. Paired 95% intervals use 5,000 question-bootstrap draws (seed 20260905), conditional on frozen fits and observed grading/mixed support.

Table 22: Matched-report intervals under strict grading. The Qwen/Llama points reproduce the original predictions; the bootstrap stream here is distinct from their frozen primary analysis. Llama's contrast upper endpoint rounds to negative zero (approximately -6.5×10^{-6}), and its sign is not robust to the format sensitivity below.

Model	KV24 [95% CI]	KV24+LM5 minus LM5 [95% CI]
Qwen3-8B	.4721 [.4380,.5060]	−.0505 [−.0812,−.0195]
Llama3.1-8B	.5157 [.4838,.5467]	−.0194 [−.0390,−.0000]
Granite-3.3-8B	.4856 [.4474,.5225]	−.0133 [−.0440,.0162]

Granite's feature validity substantially exceeds its strict scoring coverage: 3,291 attempts have valid features, but only 2,240/4,000 are scorable. JSON and multiple-choice format failures account for the gap. Invalid attempts remain in attempted denominators and are excluded from ranking; format rejection does not imply an incorrect content answer.

Table 23: Granite on the original cross-domain questions, under strict grading. Correct counts condition on scorable responses. Science and instruction dominate the overall mixed-question average; code has only five mixed questions.

Domain	Scorable/all	Correct	Mixed q	KV24	Hidden24	LM5	KV24+ LM5
Science	1,229/1,600	715	148	.5208	.6067	.5818	.5509
Instruction	783/1,200	222	85	.4234	.5670	.3615	.3684
Code	228/1,200	7	5	.5000	.5000	.5000	.6667

Granite's overall comparison favors Hidden24, while the code row has too little mixed support to resolve KV24 ranking (Table 23). The earlier 300-question extension also used different questions and a multiple-choice code interface rather than JSON. Its stronger code result therefore cannot identify a model-only effect.

Common mixed-question sensitivity. Only 99 questions are mixed for all three models, including just two code questions. Resampling these identities jointly fixes the question set but still allows different scorable-response counts. The resulting shifts in Table 24 reflect outcome-selected conditioning, not improved generation or independent evidence.

Display-wrapper sensitivity. Some feature-valid outputs are rejected solely for wrapping JSON in backticks. We remove one matching outer inline backtick pair or a complete outer triple-backtick block, optionally labeled `json`, identically for all models. The rule edits no answer content, extracts nothing from prose, and cannot recover missing boundaries or letter-plus-option-text answers. Strict scoring remains the main analysis.

Table 24: Descriptive concordance on the 99 common mixed questions under strict grading. Δ is KV24+LM5 minus LM5. This outcome-selected support does not replace each model's all-domain result.

Model	KV24	Hidden24	LM5	KV24+ LM5	Δ [95% CI]
Qwen3-8B	.6150	.6598	.6533	.6867	.0334 [−.0108,.0784]
Llama3.1-8B	.5370	.5597	.5696	.5643	−.0053 [−.0399,.0281]
Granite-3.3-8B	.4959	.6398	.5555	.5105	−.0450 [−.0908,−.0013]

For Granite, this recovers 654 scorable responses, mostly code. Code mixed support rises from 5 to 34 questions, but the KV increment remains unresolved: .0679 [−.0363,.1772]. Overall results appear in Table 25.

Table 25: Granite's post-hoc display-wrapper sensitivity, with 4,000 attempts retained in both denominators. Strict scoring is the main analysis of this descriptive extension; unwrapping changes both valid and mixed support.

Grading	Scorable	Mixed q	KV24	Hidden24	LM5	KV24+ LM5	Δ [95% CI]
Strict	2,240	238	.4856	.5903	.5014	.4881	−.0133 [−.0440,.0162]
Unwrapped	2,894	270	.5031	.5954	.5173	.5108	−.0064 [−.0354,.0221]

Under the same rule, Qwen's KV increment remains negative, whereas Llama's interval now crosses zero. Granite remains below Hidden24, with no resolved overall increment over LM5. Thus formatting changes support and some interval signs, but does not establish a positive overall KV increment in any model.

D.4 Additional source and target evaluations

Matched source-only replication. The Llama replication reuses the Qwen 800-question GSM8K cohort, 480/160/160 question split, eight samples per question, and answer-boundary endpoint, with Llama-relative layers 27–29. All 6,400 attempts remain in accounting: 5,922 are valid natural stops and the rest are invalid-format or missing-boundary rejects. Exact-token replay reproduces KV, LM, and hidden features without recorded drift. This replication reuses GSM8K outcomes already inspected in the Qwen study and is therefore descriptive.

Table 26: Matched Llama3.1-8B source-only replication on 1,197 valid test attempts from 160 questions (162 wrong). The prespecified KV24 gate—the 5,000-draw interval lower bound exceeding .5—passes. Hidden24 is a train-label-selected, label-advantaged control; comparator intervals are conditional diagnostics, not multiplicity-adjusted superiority tests.

Method	Test AUC	Question-clustered 95% CI	Accuracy at 70% coverage
KV24	.7724	[.7144,.8284]	.9262
Hidden24	.8193	[.7489,.8845]	.9481
LM5	.8154	[.7560,.8675]	.9455
KV24+LM5	.8467	[.7972,.8914]	.9469
Hidden4096	.7918	[.7332,.8435]	.9359

The source result establishes answer-boundary KV readability in a second model on matched math questions (Table 26).

Zero-target-fit MathQA transport. GSM8K coordinate choices, scalers, readouts, and cutoffs remain frozen for MathQA. Of 1,900 attempts, 1,521 are eligible (1,220 correct and 301 wrong); invalid or capped attempts count as rejects and do not enter AUC. Exact-token replay passes before

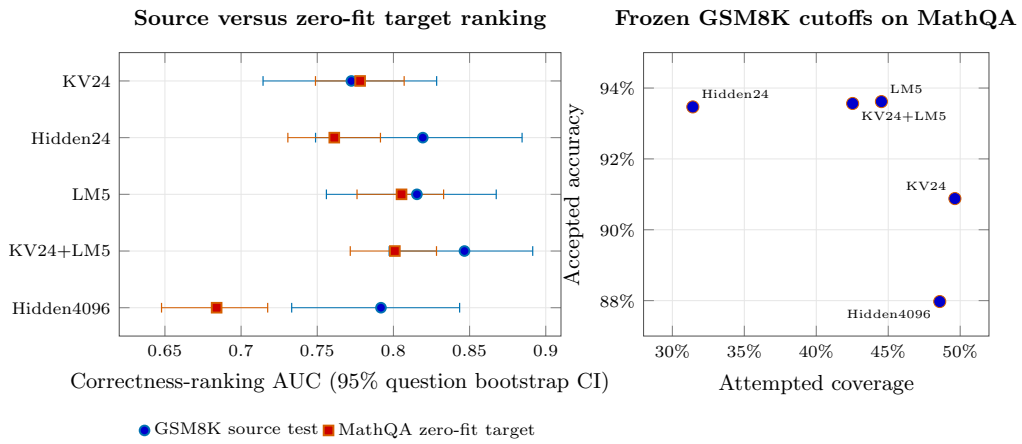

Figure 3: Matched Llama3.1-8B source and zero-target-fit MathQA evaluation. Left: source and target AUC with 5,000-draw question-bootstrap intervals. Right: attempted coverage and accepted accuracy after transporting each frozen GSM8K raw-score cutoff to MathQA. Target labels fit nothing; the selective points are reported on the fixed target cohort.

the first gold join for this run. Because MathQA outcomes had already been inspected elsewhere in the project, this transport analysis is retrospective.

KV24 ranking transfers to MathQA (Figure 3). Its paired difference from Hidden24 is unresolved, and no advantage over LM5 is established. Source cutoffs also retain a selective operating point, but without target-risk certification.

Outcome-isolated Granite extension on a separate target. The same KV24 and LM5 operators are applied to Granite-3.3-8B, with Hidden24 selected using GSM8K source-training labels only. On the 160-question source test, KV24 passes the readability gate while remaining below the control point estimates.

A separate target protocol then freezes 300 new questions, eight responses per question, and equal-weight mixed-question concordance before the private label join. It retains 2,271/2,400 scorable responses and 242 mixed questions. KV24 passes the aggregate endpoint at .5620 with a one-sided lower bound of .5330. However, the instruction-domain direction reverses, and paired comparisons favor Hidden24 and LM5. Because these questions were developed for Granite, this positive extension does not isolate a model effect or establish a domain-invariant direction.

D.5 Answer-interface sensitivity

Two supporting GSM8K studies test whether changing the answer interface or aligning its score distribution preserves performance. They concern Qwen interface shifts, rather than the cause of the cross-model differences above. Complete condition matrices and protocol decisions are retained in the supplementary materials.

Fixed marker replacement. The first study branches each pre-marker cache through its native marker and fixed ASCII `Final answer:` , then applies the frozen source KV24 readout. Four conditions cross the native `Final answer:` and `Answer:` interfaces with temperatures .9 and .3. The paired test has 1,980 common responses per condition from 498 questions; intervals use 5,000 question resamples. Canonicalization yields no resolved AUC improvement in any condition and reduces pooled accuracy at 70% coverage by 2.15 percentage points [1.21,3.15]. Thus local marker replacement does not repair the observed interface dependence. The held-out reserve is not used.

Distribution alignment. The second study uses four naturally generated interfaces. A 200-question unlabeled split estimates condition-wise mean/std or empirical-CDF maps; a separate 100-question calibration split fits shared calibration and thresholds. The common test contains 498 questions and

1,909 responses per interface. Mean/std alignment preserves within-interface ranking and raises KV24's cross-interface AUC from .6865 to .7297; its paired gain is .0432 [.0209,.0636]. The gain is relative to unaligned KV24; aligned Hidden4096 remains competitive. The prespecified analysis fails its validity-loss limit (3.031 points versus 2) and bootstrap-support requirement (7,367 valid draws out of 10,000 versus 9,500 required). These conditional estimates do not support a general interface repair.

E Answer Selection and Validation Uncertainty

E.1 Selection protocol and validation choices

This analysis selects among previously source-trained KV24, Hidden24, LM5, and KV24+LM5 readouts, keeping coefficients and directions fixed without target fitting.

Selection and evaluation. Each model has 160 source-validation and 160 source-test questions, with eight attempts per question. The AUC-selected rule maximizes validation pooled AUC. The Best-of-8-selected rule maximizes validation Best-of-8 success, breaking ties by validation AUC; remaining ties in either rule follow KV24, Hidden24, LM5, KV24+LM5 order. The selected methods are then held fixed on source test, the original 500-question cross-domain suite under strict grading, and OpenBookQA for Qwen/Llama. OpenBookQA uses sample indices 0–7 from its original 16-response pools, selected before validity or correctness filtering. It is a new eight-attempt diagnostic, not a replacement for the original 16-response concordance endpoint.

For each question, Best-of-8 is the correctness of a highest-scoring eligible response. Exact score ties are resolved uniformly over tied responses, and we report expected correctness rather than draw an extra random tie break. Majority voting selects a most frequent normalized answer, uniformly over tied answer groups. Normalization follows the existing grading convention: numeric equivalence for GSM8K, option letters for OpenBookQA, and canonical JSON or option letters for the cross-domain suite. Correctness never breaks a selection tie. Invalid attempts are unselectable but remain in the eight-attempt budget; all-invalid questions score zero and remain in the accuracy denominator. No display-wrapper repair is applied here.

Pooled AUC uses eligible responses; concordance uses mixed questions; accuracy uses all questions. Original paired 95% intervals use 5,000 question draws (seed 20260905), fixing validation choices and pairing methods with voting. The added sensitivity also resamples validation questions. Both analyses describe uncertainty conditional on the frozen fits.

Validation choices. The AUC and Best-of-8 rules choose different readouts for Qwen and Granite (Table 27). For Llama, both choose Hidden24: its Best-of-8 tie with LM5 is broken by AUC.

E.2 Answer choice on source and target pools

Changing the selection metric raises Qwen's observed cross-domain accuracy by 6.6 percentage points (Table 28), but the interval reaches zero when validation selection is repeated (Table 31). No other evaluation set or model shows a resolved benefit. On the cross-domain suite, every Best-of-8-selected readout is below same-pool voting in point accuracy, with the Qwen gap unresolved. These comparisons concern the four compact readouts; historical raw-hidden experiments use separate response pools.

Table 27: All four source readouts, with 160 questions in each split per model. Only validation results select a readout. AUC uses eligible responses; concordance averages mixed questions; Best-of-8 includes all questions. Granite's one all-invalid validation pool contributes zero success.

		Source validation			Source test		
Model	Readout	AUC	Concordance	Best-of-8	AUC	Concordance	Best-of-8
Qwen3-8B	KV24	.8145	.6481	92.50%	.9166	.6849	93.75%
	Hidden24	.8094	.8756	95.00%	.9050	.7913	94.38%
	LM5	.7245	.7385	92.50%	.7251	.6730	94.38%
	KV24+LM5	.8432	.7297	94.38%	.9217	.7192	94.38%
Llama3.1-8B	KV24	.7814	.6707	86.88%	.7724	.6756	88.12%
	Hidden24	.8522	.7005	88.12%	.8193	.7274	89.38%
	LM5	.8251	.7023	88.12%	.8154	.6629	86.88%
	KV24+LM5	.8519	.7402	85.62%	.8467	.7129	89.38%
Granite-3.3-8B	KV24	.6889	.6602	87.50%	.7509	.6262	89.38%
	Hidden24	.7706	.6150	85.62%	.8086	.6473	91.88%
	LM5	.7512	.6212	87.50%	.8326	.6104	91.25%
	KV24+LM5	.7422	.6486	88.12%	.8490	.6879	90.62%

Table 28: Best-of-8 accuracy under the AUC-selected and Best-of-8-selected readouts. Source test has 160 questions per model; each target has 500. All pools contain eight attempts. Majority uses the same eligible answers; fractional accuracies arise from expected correctness under uniform ties. Each difference cell gives the point estimate above its paired 95% interval; intervals are descriptive and conditional on the validation choices.

Model	AUC-selected accuracy	Best-of-8-selected accuracy	Selection difference, pp	Vote	Vote difference, pp
Source test					
Qwen3-8B	94.38%	94.38%	0.00 [−1.88,+1.88]	94.69%	−0.31 [−0.94,0.00]
Llama3.1-8B	89.38%	89.38%	0.00 [0.00,0.00]	93.75%	−4.38 [−8.75,−0.62]
Granite-3.3-8B	91.88%	90.62%	−1.25 [−5.00,+1.88]	94.48%	−3.85 [−7.29,−0.83]
Cross-domain					
Qwen3-8B	53.20%	59.80%	+6.60 [+2.40,+11.00]	61.55%	−1.75 [−5.68,+2.30]
Llama3.1-8B	25.40%	25.40%	0.00 [0.00,0.00]	32.32%	−6.92 [−10.47,−3.45]
Granite-3.3-8B	33.00%	30.50%	−2.50 [−6.20,+1.00]	37.45%	−6.95 [−10.31,−3.68]
OpenBookQA					
Qwen3-8B	89.80%	90.40%	+0.60 [−1.00,+2.40]	90.20%	+0.20 [−1.00,+1.40]
Llama3.1-8B	80.80%	80.80%	0.00 [0.00,0.00]	86.10%	−5.30 [−7.90,−2.80]

Table 29: All four readouts on the same target response pools, with 500 questions and eight attempts per question and model. AUC and concordance condition on eligible and mixed support, respectively; Best-of-8 includes all questions. All scores use frozen source readouts.

Dataset	Model	Readout	Pooled AUC	Macro concordance	Best-of-8
Cross-domain	Qwen3-8B	KV24	.5273	.4721	47.20%
		Hidden24	.6500	.6711	59.80%
		LM5	.6514	.6213	58.00%
		KV24+LM5	.6379	.5708	53.20%
Cross-domain	Llama3.1-8B	KV24	.6977	.5157	24.80%
		Hidden24	.7034	.5336	25.40%
		LM5	.6281	.5766	29.40%
		KV24+LM5	.7532	.5572	30.20%
Cross-domain	Granite-3.3-8B	KV24	.6365	.4856	27.90%
		Hidden24	.7151	.5903	33.00%
		LM5	.4352	.5014	30.00%
		KV24+LM5	.5665	.4881	30.50%
OpenBookQA	Qwen3-8B	KV24	.6998	.5591	89.60%
		Hidden24	.6203	.5754	90.40%
		LM5	.6579	.5576	89.00%
		KV24+LM5	.7165	.5617	89.80%
OpenBookQA	Llama3.1-8B	KV24	.5743	.5638	82.00%
		Hidden24	.5868	.5391	80.80%
		LM5	.6754	.6025	86.40%
		KV24+LM5	.6607	.6006	85.00%

E.3 Validation-selection uncertainty

Following the reselection procedure in Section 7, we resample each model's 160 validation questions 5,000 times. Question multiplicities weight both AUC and Best-of-8, with all response siblings retained. All draws contain both classes; probe fits and training transforms remain fixed.

Table 30 reports empirical selection frequencies. Table 31 adds the resulting validation-and-test percentile intervals to the original point contrasts. These describe sensitivity to the observed validation sample, not posterior method probabilities or a new confirmatory test.

Table 30: Selection frequencies across validation-question resamples. The Best-of-8 rule uses AUC to break utility ties. Each row sums to 100%; frequencies describe the observed validation sample.

Model	Selection rule	KV24	Hidden24	LM5	KV24+LM5
Qwen3-8B	AUC	2.24%	21.70%	1.52%	74.54%
	Best-of-8	0.14%	71.06%	0.02%	28.78%
Llama3.1-8B	AUC	0.00%	46.70%	8.86%	44.44%
	Best-of-8	16.58%	47.54%	35.74%	0.14%
Granite-3.3-8B	AUC	0.06%	65.08%	26.10%	8.76%
	Best-of-8	23.96%	9.30%	23.04%	43.70%

Reselection makes the Qwen cross-domain interval reach zero. Llama's original policies coincide, but can diverge on resampled validation questions, giving a nonzero-width interval. Across all evaluation sets, no model has a strictly positive lower bound for Best-of-8-selected minus AUC-selected accuracy (Table 31). The observed Qwen gain therefore leaves the benefit of changing the selection metric unresolved.

The error partition in Table 32 uses the original Best-of-8-selected readout. Per-question availability and selection errors are included in the standalone replay (Appendix G).

Table 31: Original point contrasts with descriptive 95% intervals resampling both source validation and evaluation questions. Readout fits stay fixed. The last column compares the reselected Best-of-8 policy with the same-pool vote. All questions, including all-invalid pools, remain in accuracy denominators.

Dataset	Model	Original Best-of-8 minus AUC, pp	Validation+test interval, pp	Best-of-8 minus vote interval, pp
Source test	Qwen3-8B	0.00	[−1.25,+1.25]	[−1.25,+0.62]
	Llama3.1-8B	0.00	[−5.63,+3.12]	[−10.62,−1.25]
	Granite-3.3-8B	−1.25	[−5.00,+1.88]	[−7.92,−0.73]
Cross-domain	Qwen3-8B	+6.60	[0.00,+10.80]	[−11.02,+1.87]
	Llama3.1-8B	0.00	[−7.60,+5.80]	[−10.20,−0.60]
	Granite-3.3-8B	−2.50	[−7.20,+2.10]	[−11.72,−3.16]
OpenBookQA	Qwen3-8B	+0.60	[−0.80,+2.20]	[−1.50,+1.40]
	Llama3.1-8B	0.00	[−6.40,+7.00]	[−7.50,+2.00]

Table 32: Cross-domain error accounting for the original Best-of-8-selected readout on all 500 questions per model. Availability error means no eligible correct candidate; selection error means an eligible correct candidate is available but missed. All-invalid pools contribute to availability error.

Model	Availability error	Selection error	Total error	All-invalid q.
Qwen3-8B	16.20%	24.00%	40.20%	0
Llama3.1-8B	42.20%	32.40%	74.60%	0
Granite-3.3-8B	44.20%	25.30%	69.50%	27

Earlier ambiguity-gated math selectors likewise showed no stable advantage over voting. Their candidate-count and consensus-stratified reports are retained in the supplement as separate historical analyses; they use different pools and policies from the source-selected readouts evaluated here.

F Selective Filtering, Calibration, and Risk

F.1 Filtering at matched coverage

The following comparison uses exact rank-matched coverage on the aligned external 500-question rows. Lower area under the risk–coverage curve (AURC) and excess AURC (E-AURC) are better. Accept-all accuracy is 90.84% for every method; only the ordering changes.

Table 33: Selective ranking at matched realized coverage. E-AURC subtracts the oracle ordering with the same label count. Historical comparators retain their previously released schemas and are not refit here.

Method	AURC	E-AURC	Accuracy @ 50% cov.	Accuracy @ 70% cov.	Accuracy @ 90% cov.
KV24	.02960	.02526	97.52%	97.14%	94.32%
KV+surface+LM (historical)	.02818	.02384	97.90%	96.79%	94.57%
LM-only (historical)	.04050	.03616	96.25%	94.89%	93.14%
Surface+LM (historical)	.03621	.03187	96.20%	95.18%	93.36%
Hidden+controls (historical)	.03130	.02696	97.00%	96.05%	93.99%

The supplementary operating-point and risk-decile reports describe review cost and error concentration for this same score. They provide additional views of the filtering result without changing its evaluation cohort.

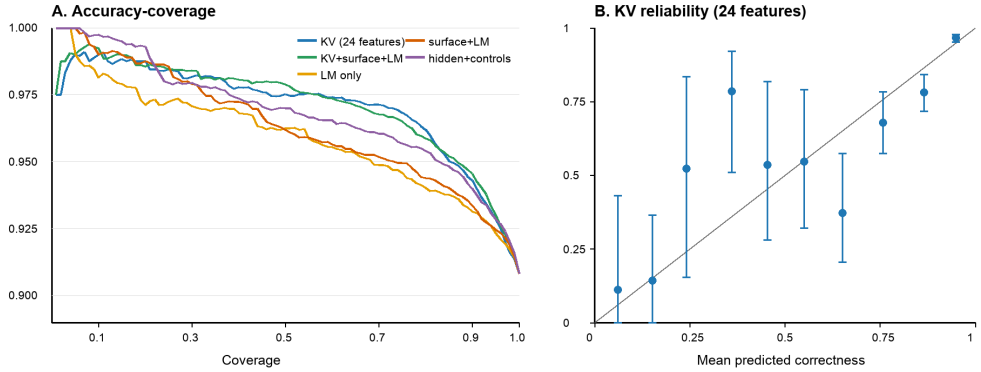

Figure 4: Frozen new-seed GSM8K selective-prediction detail. Left: exact rank-matched accuracy–coverage curves on aligned responses. Right: corrected 24-feature KV-only reliability with question-clustered 95% intervals; sparse low-confidence bins are correspondingly uncertain. Historical comparator artifacts retain their released schemas and are not refit here, so this diagnostic supports operational filtering rather than KV specificity.

F.2 MathQA probability and threshold transfer

For the historical Qwen MathQA comparison, the source probe is frozen and a positive-slope probability map and threshold are fit on target development. The threshold uses 244 eligible rows from a prespecified 250-question calibration subset. Table 34 reports the historical write-once test, reanalyzed retrospectively here. Ranking retains the source score direction; probability calibration and threshold fitting use target labels.

Table 34: Complete historical write-once MathQA comparison on 1,847 strict common first responses, now reported as a reanalysis. Each probability map and threshold was frozen on target development; AUC uses the unchanged source ranking direction.

Locked MathQA method	AUC [95% CI]	Brier	ECE (10 bins)	Fixed coverage	Accepted accuracy
KV24	.6597 [.6169,.7008]	.1103	.0532	70.06%	90.96%
LM5	.7864 [.7553,.8156]	.1009	.0526	75.04%	92.35%
Hidden4096	**.7898** [.7595,.8186]	.1015	.0292	74.23%	**93.22%**
KV24+LM5	.7518 [.7132,.7883]	**.0990**	**.0210**	74.12%	93.06%

Adding KV24 to LM5 reduces pooled AUC by .0346 [.0060,.0633] in the paired comparison. Improved probability calibration of the combined readout therefore does not imply improved discrimination.

F.3 Probability calibration versus score-only coverage

Positive-slope adapters are fit on the fixed 244-question calibration set and evaluated on a disjoint 603-question clean holdout, which excludes 31 earlier engineering-pilot rows. Table 35 shows improved probability calibration with target labels. In contrast, raw, offset-only, and positive-Platt logits are affine monotone coordinates and their exact-coverage masks agree row by row at every reported coverage under the frozen tie-breaking rule.

Outcome-free target quantiles instead fix acceptance masks before labels are joined. The resulting KV rule has empirical risk below the 10% target, but its one-sided Clopper–Pearson upper bound exceeds that target (Table 37). This conditional binomial diagnostic fixes the rule and does not cover quantile estimation, selection, or deployment shift.

Table 35: Probability calibration on the disjoint 603-question clean holdout. The label-fitted columns are not zero-label methods.

Method	Raw Brier	Offset Brier	Platt Brier	Raw NLL	Offset NLL	Platt NLL
KV24	.1732	.1021	.1017	.5277	.3722	.3652
LM5	.2538	.1375	.1016	1.3867	.7529	.3425
Hidden4096	.1248	.1015	.1007	.4015	.3425	.3400
KV24+LM5	.1496	.0999	.0950	.5211	.3824	.3316

F.4 Learn-then-test sensitivity

We apply fixed-sequence learn-then-test (LTT) to the opened calibration artifacts. For each method, the 50/70/90% candidate thresholds are the outcome-free 244-question calibration-score quantiles and are therefore fixed independently of the disjoint 603-question holdout labels. At a candidate with m accepted rows and e errors, we test the null that conditional accepted risk exceeds .10 using the exact lower-tail boundary p-value $\Pr\{\mathrm{Binomial}(m, .10) \leq e\}$. The sequence proceeds from lower to higher requested coverage and stops at the first non-rejection, at level $\delta = .05$ within each method.

Table 36: Fixed-sequence LTT baseline on the 603-question clean holdout. Method-specific sequences are separate 5% testing families; the $\delta/4$ column is a Bonferroni-across-four-methods sensitivity. Because the procedure was specified after the holdout outcomes were open, results are retrospective. "No policy selected" means the sequence stopped at its first non-rejection.

Method	Method-specific selected coverage	$\delta/4$ coverage	Selected 95% CP UCB
KV24	No policy selected		
LM5	50%	50%	.0778
Hidden4096	70%	50%	.0960
KV24+LM5	70%	70%	.0880

KV24 stops at its first candidate, while the other readouts admit operating points (Table 36). Its empirical risk remains below the target, but uncertainty prevents rejection of the risk null.

F.5 All-attempt and full-procedure uncertainty

The historical cohort manifest conserves all 1,900 attempts: 1,847 strict-score-eligible, 10 malformed-marker, and 43 missing-boundary. Hard rejection yields the four rows in Table 37; invalid attempts never enter AUC or the accepted-set binomial count.

Table 37: MathQA all-attempt sensitivity. All rules use score-only frozen masks; 53 of 1,900 attempts lack a strict boundary and are hard-rejected. The exact 95% bound is a conditional one-sided Clopper–Pearson bound for the frozen acceptance rule.

Rule	Accepted	Errors	Eligible cov.	Attempted cov.	Exact 95% UCB risk
KV24	1,294	117	70.06%	68.11%	10.46%
LM5	1,386	106	75.04%	72.95%	8.93%
Hidden4096	1,371	93	74.23%	72.16%	8.01%
KV24+LM5	1,369	95	74.12%	72.05%	8.18%

Refitting the source and target stages widens KV24's uncertainty, and its risk interval still extends above the target (Table 38). This pipeline has its own point fit and acceptance mask; it is separate from the designated frozen rule in Table 37.

The paired draw plan resamples all sibling source rollouts as question clusters and shares evaluation and target-primary draws across methods. Hidden24 selection and PCA are repeated inside every source-training draw. The 20 frozen random-feature controls are not present in this tensor.

Table 38: Five-hundred-draw question-clustered uncertainty for the separate refit pipeline. "Fixed model" fixes that pipeline's point model and resamples evaluation questions; it is not the historical write-once rule. Full procedure refits source transforms/readouts/calibration and the target adapter/quantile. All 500 draws succeed with no fallback.

Bootstrap level	Method	Mean risk	95% interval	Mean coverage
Fixed model	KV24	.0905	[.0755,.1069]	.6900
Target refit	KV24	.0908	[.0756,.1072]	.6936
Full procedure	KV24	.0925	[.0684,.1154]	.6969
Full procedure	Hidden24	.0639	[.0452,.0916]	.7092
Full procedure	PCA24	.0601	[.0442,.0777]	.6967
Full procedure	LM5	.0763	[.0577,.0959]	.7480

G Released Artifacts and Reproducibility

The anonymous bundle contains aggregate reports, numeric provenance, and figure assets. The cross-domain and verifier-selection bundles provide row-level statistical replays, including Figure 1's source decomposition. The supplementary documentation maps each analysis to its data, reports, and execution commands. Local usernames, cluster paths, and scheduler logs are excluded.

The release supports these replays and transcription checks. It does not include every cohort's full generation, extraction, fitting, and calibration pipeline, and therefore does not reproduce the whole study end to end.

Standalone statistical replay. The verifier selection supplement supplies 27,680 normalized response records and the same analysis script. With NumPy, one command recomputes the fixed-choice comparisons, validation selection frequencies, reselection intervals, and per-question error partitions from frozen scores. The reference comparison covers every scientific field in the released aggregate report. This reproduces the statistical analysis; it starts after raw-text grading, feature extraction, and source readout fitting.

Additional exploratory analyses. The supplement retains the earlier write-ups for head localization, trajectory/suffix features, state perturbations, target-supervised geometry, historical selectors, and interface repairs, alongside the released aggregate reports. Head localization and trajectory controls did not establish the tested KV advantage; generic perturbations also moved the readouts; target-label adaptation depended on target supervision. These results support neither a localized causal correctness mechanism nor a universal source direction. Superseded Llama feasibility attempts produced no cross-model performance result and are kept only as provenance.