QUERY–KEY ENSEMBLES FOR REASONING VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models typically generate reasoning chains that are then judged from the outside, by solvers, reranking heuristics, or trained reward models. We ask whether the model has already formed a judgment of its own. Given a generated solution, we read which of two explicit verdicts, correct or erroneous, the model's attention heads prefer, using raw query–key (QK) inner products taken before Q/K normalization, rotary position encoding, and the softmax. Building on QK readouts for multiple-choice selection and logical consistency, we turn these head-wise preferences into a verifier of complete chain-of-thought (CoT) solutions without updating the model. A calibration-free rule, CF-QK, weights each head by its agreement across read positions and its responsiveness across an unlabeled batch; a linear readout, L2-QK, needs only 50 labeled solutions. Across Qwen3 (8B–32B) and DeepSeek-R1-Distill-Llama-8B on MATH-500 and HLE-1/4, these ensembles outperform the model's own verdict logits, solution likelihood, and prompted self-verification in most settings. Without labels, CF-QK raises balanced accuracy over verdict logits by up to 18 points. Across ten 50-label calibration splits, L2-QK gains 4–16 points over calibrated logits on every model and dataset. The fitted rules transfer without target labels to GSM8K and SVAMP, and each verdict costs one forward pass. Matched edits to solution content and activation patching show that the scores respond to the reasoning itself and that the selected heads shape the model's verdict. Complementary experiments study multiple-choice selection with or without a "think-first" CoT preface.

1 INTRODUCTION

Evaluating generated mathematical solutions is necessary both for selecting answers and for identifying errors. Chain-of-thought prompting and self-consistency improve generation and answer aggregation (Wei et al., 2022; Wang et al., 2023), while learned verifiers and process reward models assess candidate solutions (Cobbe et al., 2021; Qwen Team, 2025). The generating model's internal representations provide another source of information: a correct answer may be represented internally even when it is not preferred by the output logits.

Query–key (QK) alignment measures interactions between query and key representations before attention softmax. Previous work has used these interactions for multiple-choice selection and logical-consistency assessment (Tulchinskii et al., 2024; 2025). In multiple-choice tasks, individual heads can identify correct options that decoding misses. We retain the QK inner product and ask how to turn head-wise preferences into a decision about an already-generated mathematical solution. We examine head behavior across models, extraction points, prompting conditions, and datasets.

1.1 CONTRIBUTIONS

We make three contributions.

Calibration-free verification. We introduce CF-QK, which aggregates semantic correct-versus-error margins using agreement across read positions and responsiveness across an unlabeled reference batch. It requires neither correctness labels nor a reference answer at scoring time, and uses one pass through a frozen model.

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.

Verification with a small calibration set. We compare single-head selection, top-K averaging, an L2-regularized linear classifier (L2-QK), and subset search on the same QK features, with separate search, validation, and threshold-fitting data. Across ten 50-label calibration splits per model, L2-QK has higher mean balanced accuracy (BAcc) than calibrated verdict logits on MATH and HLE-1/4 for all four models. Fitted readouts are held fixed for transfer.

Analysis of the verification signal. We evaluate aggregation components and independent source reference batches across four models, and examine reference-batch variation, calibration sensitivity, controlled changes in solution content, and activation interventions. Multiple-choice selection provides complementary evidence about individual QK heads.

2 Related work

Reasoning, verification, and answer selection. Chain-of-thought prompting elicits intermediate reasoning (Wei et al., 2022), and self-consistency selects an answer by aggregating sampled reasoning paths (Wang et al., 2023). Outcome verifiers (Cobbe et al., 2021) and process reward models (Qwen Team, 2025; Lightman et al., 2023) learn to evaluate solutions or their intermediate steps. Verbalized confidence (Lin et al., 2022), sequence likelihood, and self-certainty (Kang et al., 2025) provide other measures of response quality. Prompted self-verification is a weak judge: models rarely locate their own reasoning errors without external feedback (Huang et al., 2024; Tyen et al., 2024). We evaluate candidate solutions using features extracted from a frozen model's attention projections, with either unlabeled aggregation or a fitted decision rule.

Internal representations. Contrast-consistent search extracts information from model activations using consistency constraints without correctness labels (Burns et al., 2022), although unsupervised criteria can track salient features unrelated to truth (Farquhar et al., 2023). Probes on hidden states predict the correctness of answers and reasoning steps (Orgad et al., 2025; Zhang et al., 2025a; Zhao et al., 2026), but they are fitted on in-domain labels, and error detectors often fail to transfer across datasets (Orgad et al., 2025). Attention statistics of a frozen model also serve as correctness or hallucination signals: attention mass on constraint or context tokens (Yuksekgonul et al., 2024; Chuang et al., 2024), per-head attention entropy (Ostmeier et al., 2026), and spectral features of attention graphs (Noël, 2026); attention weights further yield zero-shot rerankers (Chen et al., 2025). These statistics summarize where a model attends, whereas we read a signed preference between two explicit verdicts.

QK scores. Select-and-copy heads provide answer-selection signals in multiple-choice tasks (Tulchinskii et al., 2024), and QK alignment has been used to classify the validity of controlled logical inferences (Tulchinskii et al., 2025). Both readouts use a single selected head and judge material supplied in the prompt: fixed answer options or a premise–statement pair. We instead study solution-level verification of complete reasoning traces generated by the model itself, where the verifier must integrate information across a heterogeneous chain of thought and produce a single correctness judgment. We apply the same QK primitive to this task, score a contrast between two explicit verdict options, and replace the single head with ensembles that use no correctness labels (CF-QK) or 50 labels (L2-QK). Appendix G lists the differences from these readouts.

3 Internal preferences in multiple-choice reasoning

We examine individual QK scores in multiple-choice answering before and after deliberation. Full protocols and results appear in Appendix E; Figure 7 in Appendix F illustrates the query and key positions.

Answer selection before and after deliberation. For multiple-choice question answering (MCQA), the answer-boundary query is compared with keys at the ends of the options, directly after the prompt or after a chain of thought of at most 1,200 generated tokens. A head selected on 500 calibration questions is evaluated on 2,000 disjoint MMLU-Pro questions across five splits. Permutation accuracy counts joint correctness before and after option shuffling (Gupta et al., 2024), with calibration repeated for the permuted split.

On LLaMA-3.1-8B, plain QK selection reaches 32.9% accuracy versus 16.4% for the output baseline, the model's own answer prediction. After deliberation, the baseline reaches 31.4%, close to QK's

32.1%. On Qwen3-8B, the native-normalized readout reaches 54.7% after CoT versus 53.9% for the baseline, although its permutation accuracy is slightly lower. The full results on MMLU-Pro and on HLE-MC, a 2,100-question multiple-choice adaptation of Humanity's Last Exam (Phan et al., 2025), appear in Appendices E.2 and E.3. These comparisons characterize option selection across models, extraction conventions, and prompting conditions.

From a useful head to an aggregation rule. In verification, all heads compare the same two verdicts, which provides a common score for aggregation. CF-QK estimates weights from unlabeled examples. With source labels, we compare head selection, subset search, and a regularized linear classifier. We evaluate these methods on candidate solutions separately from the multiple-choice experiments.

4 Query–key scores and ensemble verification

4.1 Query–key scores

Let $q_h(t)$ and $k_h(t)$ denote the query and key vectors at position t for head h. We measure their unscaled inner product,

$$s_h(t_q, t_k) = q_h(t_q)^\top k_h(t_k). \tag{1}$$

For multiple-choice selection, the answer-boundary query is compared with keys at the ends of the options. Verification presents a problem and candidate solution, without the reference answer, followed by two options stating that the candidate is correct or contains an error, and then `Verdict:`. Its semantic margin is

$$d_{h,m,v}(x) = q_h(t_q)^\top [k_h(t_c) - k_h(t_e)], \tag{2}$$

where m specifies the option mapping and v the query and key positions. Positive margins favor correctness. The first mapping assigns correct to A and error to B; the second reverses them.

Verification reads raw query/key projections before native Q/K normalization and rotary encoding, without softmax or dimension scaling. With grouped-query attention, query head h uses key head $\lfloor h/g \rfloor$, where g is the sharing factor. Removing rotary encoding from the dot product does not remove position information from contextual representations. Appendix A gives token anchors and extraction details; Appendix E.1 specifies the separate MCQA convention.

4.2 Calibration-free aggregation

CF-QK combines agreement across read positions with variation across an unlabeled batch $\mathcal{B}$. Eligible heads are in the final half of the layers. Three query anchors in `Verdict:` and three key anchors at each option's end give nine nominal views; duplicate measurements are removed. Each head casts the mean vote

$$u_h(x) = \frac{1}{|\mathcal{V}(x)|} \sum_{v \in \mathcal{V}(x)} \operatorname{sign}(d_{h,0,v}(x)). \tag{3}$$

Let c_h be the fraction of examples with $|u_h| > 10^{-8}$, a_h the batch mean of $|u_h|$, and p_h the fraction of non-tied votes favoring correctness. The weight is

$$w_h = \mathbf{1}[\ell(h) \geq \lfloor L/2 \rfloor]\, 4p_h(1-p_h)c_h a_h. \tag{4}$$

Agreement and coverage favor consistent nonzero votes, while $4p_h(1-p_h)$ suppresses heads that always predict the same verdict. These weights are heuristics computed without labels and do not estimate head accuracy. The one-pass score is

$$S_{\mathrm{CF}}(x) = \frac{\sum_h w_h u_h(x)}{\sum_h w_h}, \qquad \widehat{y}(x) = \mathbf{1}[S_{\mathrm{CF}}(x) > 0]. \tag{5}$$

If all weights vanish, we assign equal weights to the eligible heads. CF-QK-R1P uses no labeled calibration; its score depends on the reference batch, which includes the scored examples. The two-pass CF-QK-RSym variant averages the first-mapping score S_1 and a score S_{12} pooling both mappings:

$$S_{\mathrm{RSym}}(x) = \tfrac{1}{2}[S_1(x) + S_{12}(x)]. \tag{6}$$

Weights are recomputed for each component. This variant retains a first-mapping component and is not exactly order-invariant. We use CF-QK to denote R1P unless otherwise specified; Appendix B.3 gives the two-pass evaluation.

Design choices. The final-half layer pool is a fixed architectural choice motivated by evidence of layer-specific reasoning circuits and later-layer multi-hop computation (Hong et al., 2025; Biran et al., 2024). Read anchors follow the verdict and option boundaries; responsiveness suppresses constant verdict preferences. No correctness labels, including those of the reported evaluation examples, were used to choose the layer pool, anchors, weighting formula, or two-pass blend. Appendix A.3 details these choices and distinguishes fixed defaults from weights computed on unlabeled reference batches.

4.3 CALIBRATED QK METHODS

All calibrated QK methods use the colon-to-end-of-line view with A=correct. Scales $b_h = \mathrm{IQR}_{\mathrm{search}}(d_h) + 10^{-8}$ are fitted on search data. The candidate pool contains the 64 heads with the highest search AUROC and 64 additional heads selected to have low maximum absolute correlation with those already in the pool. For an equal-weight subset H,

$$S_H(x) = \frac{1}{|H|} \sum_{h \in H} \frac{d_h(x)}{b_h}, \qquad J(H) = \mathrm{AUROC}_{\mathrm{search}}(S_H, y). \tag{7}$$

Single head and top-K mean. The single-head rule selects one pool head by validation AUROC. Top-K averages the strongest heads ranked on the search split, with K chosen on validation data.

Linear QK (L2). An L2-regularized logistic classifier learns weights on the same normalized head features. We fit the classifier on the search split and select regularization on the validation split. Each head can receive a positive or negative weight. Appendix C.7 decomposes a fitted score into individual head contributions and the intercept.

Subset search. We also compare greedy and random subsets with SA-QK, which uses simulated annealing (Kirkpatrick et al., 1983). All three use the same feature pool, subset sizes, and objective-evaluation budget; validation AUROC selects among the search candidates. Search schedules and budgets are specified in Appendix A.

All methods use disjoint search, validation, and threshold-fitting partitions. A separate threshold maximizes source balanced accuracy; head indices, scales, fitted weights, and thresholds are frozen for transfer. Selection operates on cached scores without additional model passes. The main experiment allocates 30/10/10 examples to these partitions; the 400-label comparison uses 240/80/80 (Appendix A). Appendix A gives extraction and calibration details.

5 EXPERIMENTAL SETUP

Verification tasks and models. The primary study evaluates Qwen3-8B, Qwen3-14B, Qwen3-32B, and DeepSeek-R1-Distill-Llama-8B (DeepSeek-Llama-8B below) on MATH-500 (Hendrycks et al., 2021) and HLE-1/4 (HLE below), a deterministically constructed, text-only, four-option adaptation of Humanity's Last Exam (Phan et al., 2025). We evaluate transfer on GSM8K (Cobbe et al., 2021) and SVAMP (Patel et al., 2021). The complementary multiple-choice study in Section 3 spans LLaMA-3.1, Qwen3, and DeepSeek-R1-Distill-Qwen models from 1.5B to 32B. It uses task-specific prompts and QK read positions (Table 4).

Calibration, adaptation, and labels. Calibrated verification uses 50 source MATH labels: 30 examples for fitting or head search, 10 for validation, and 10 for threshold fitting. We repeat calibration over ten stratified splits and three search seeds per split for each model. The remaining 450 MATH examples are excluded from fitting and shared by the compared methods within each split. HLE uses 2,156 fixed examples without target-label calibration. CF-QK uses no correctness labels; its comparison and weighting ablations use the original fixed evaluation rows (Appendix A).

MATH correctness is determined by final-answer equivalence using a Qwen3-32B judge; HLE-1/4 uses exact option matching. We examine intermediate consistency separately through controlled perturbations. The broader MCQA study uses a separate 2,100-question HLE adaptation, HLE-MC, divided into 500 calibration and 1,600 evaluation questions.

Baselines and metrics. The open-generation LLM-as-judge baseline counts unparsed DeepSeek verdicts as incorrect; Qwen results use the midpoint of the lower and upper BAcc bounds (Appendix A.6). Direct verdict logits use the semantic A/B margin from the same prompt passes as QK.

Indirect prompted confidence (Lin et al., 2022) and mean solution log probability provide additional scalar scores (Appendix D.2). In the main experiment, scalar baselines use the same 50-label source calibration budget as the QK readouts. Their fitted thresholds transfer unchanged; zero-shot variants use no calibration labels. Verification uses a 4,096-token context. If necessary, we remove tokens from the middle of the candidate solution while preserving the problem and verdict options. The PRM-min comparison in Appendix B.8 uses Qwen2.5-Math-PRM-7B under a 2,048-token, 32-step budget; all methods in that comparison score the same candidate solutions.

We report balanced accuracy $\mathrm{BAcc} = (\mathrm{TPR} + \mathrm{TNR})/2$ in the verification tables. BAcc gives equal weight to the two classes at the specified threshold. Balanced accuracy is particularly relevant given HLE's class imbalance. Table 2 reports SD across ten calibration splits after averaging the three seeds within each split. For non-judge methods, Table 1 and the full fixed-split comparison in Appendix B.1 report conditional bootstrap SD from 10,000 resamples stratified by correctness, with readouts and reference weights held fixed. Appendix C.6 gives a separate Qwen3-8B label-budget study.

6 Verification results

6.1 Calibration-free verification

Table 1 compares CF-QK-R1P with direct and indirect verdict logits without calibration labels. On the fixed MATH evaluation set, CF-QK has the highest BAcc for every model, exceeding verdict logits by 8.72–17.85 points. On HLE, it is best for three of four models; for Qwen3-14B, verdict logits are 5.69 points higher. Because QK and direct verdict logits share the verification prompt, the comparison isolates the readout. Appendix B.1 retains the full fixed-split comparison, including the open-generation judge and all calibrated methods.

Table 1: Verification without calibration labels: BAcc (%) $\pm$ conditional bootstrap SD. MATH uses the original fixed 450-example evaluation set; HLE uses 2,156 fixed examples. CF-QK weights use the full unlabeled MATH or HLE batch. Best and second-best estimates in each row are in bold and underlined, respectively.

Model	Dataset	Verdict logits	Indirect logits	CF-QK-R1P
Qwen3-8B	MATH	<u>65.74</u> $\pm$ 1.71	65.60 $\pm$ 1.74	**79.36** $\pm$ 1.93
	HLE	75.61 $\pm$ 2.27	<u>77.56</u> $\pm$ 2.63	**80.84** $\pm$ 2.60
Qwen3-14B	MATH	<u>75.50</u> $\pm$ 1.97	68.80 $\pm$ 1.89	**84.22** $\pm$ 1.79
	HLE	**80.48** $\pm$ 1.96	<u>77.66</u> $\pm$ 2.02	74.79 $\pm$ 2.21
Qwen3-32B	MATH	<u>65.26</u> $\pm$ 1.81	60.00 $\pm$ 1.56	**83.11** $\pm$ 1.76
	HLE	<u>74.81</u> $\pm$ 1.63	71.67 $\pm$ 1.71	**77.91** $\pm$ 1.87
DeepSeek-Llama-8B	MATH	<u>63.55</u> $\pm$ 1.64	58.57 $\pm$ 1.30	**75.41** $\pm$ 1.95
	HLE	<u>68.26</u> $\pm$ 1.44	59.90 $\pm$ 2.15	**74.98** $\pm$ 2.29

6.2 Repeated verification with 50 calibration labels

We compare calibrated verdict logits, top-K QK averaging, and L2-QK across ten stratified 50-label splits and three search seeds per split for each of four models. Table 2 reports mean BAcc and SD over the ten splits after averaging seeds within each split. Each MATH evaluation uses the remaining 450 examples, shared by the methods within that split. HLE uses the same 2,156 examples throughout, with all fitted readouts and thresholds transferred from MATH.

L2-QK has the highest mean BAcc among the three calibrated methods in all eight model–dataset comparisons. Its paired mean improvement over calibrated verdict logits ranges from 4.06 to 16.05 percentage points, with positive differences in eight to ten of the ten splits per comparison. Top-K averaging has a higher mean than calibrated logits in seven of eight comparisons. Appendix C.5 gives paired differences, split-level wins, and threshold variation; it also shows that the fixed-split Qwen3-32B fit (Appendix B.1) uses a lower source-calibrated decision threshold than the repeated fits and examines the resulting sensitivity.

The fixed HLE population also permits comparison with the uncalibrated verdict logits in Table 1. Mean L2-QK BAcc is higher for Qwen3-8B (79.50% versus 75.61%), Qwen3-32B (79.80% versus

Table 2: Repeated 50-label calibration: BAcc (%), mean ± SD over ten stratified splits after averaging three search seeds within each split. MATH uses the split-specific 450-example evaluation set; HLE remains fixed. Best and second-best means in each row are in bold and underlined, respectively.

Model	Dataset	Calibrated logits	Top-K QK	L2-QK
Qwen3-8B	MATH	72.92 ± 2.77	$\underline{77.90} \pm 3.61$	$\mathbf{79.32} \pm 3.40$
	HLE	63.45 ± 10.66	$\underline{78.09} \pm 2.79$	$\mathbf{79.50} \pm 1.11$
Qwen3-14B	MATH	78.99 ± 2.08	$\underline{81.07} \pm 4.48$	$\mathbf{83.65} \pm 3.33$
	HLE	69.60 ± 9.57	$\underline{71.15} \pm 8.24$	$\mathbf{74.99} \pm 6.08$
Qwen3-32B	MATH	79.00 ± 2.75	$\underline{81.46} \pm 6.76$	$\mathbf{88.06} \pm 0.93$
	HLE	$\underline{74.81} \pm 6.33$	72.44 ± 9.11	$\mathbf{79.80} \pm 1.00$
DeepSeek-Llama-8B	MATH	79.06 ± 1.94	$\underline{80.69} \pm 5.36$	$\mathbf{83.12} \pm 3.22$
	HLE	63.63 ± 6.08	$\underline{71.60} \pm 13.90$	$\mathbf{79.38} \pm 6.59$

74.81%), and DeepSeek-Llama-8B (79.38% versus 68.26%). For Qwen3-14B, uncalibrated logits reach 80.48%, compared with 74.99% for L2-QK. This distinguishes improvement over source-calibrated thresholds from improvement over the original verdict rule.

6.3 Transfer across domains and model families

Table 3 reports transfer from MATH to GSM8K and SVAMP, together with DeepSeek-Llama-8B and Qwen3-32B verification results under the 400-label protocol. For CF-QK, transfer means applying the same aggregation rule with unlabeled reference statistics from the target; the fitted rules keep their heads, scales, and thresholds frozen. Without target labels, CF-QK exceeds direct verdict logits on both targets for both Qwen3 sizes. The frozen Qwen3-8B SA ensemble obtains 69.34% GSM8K and 72.75% SVAMP BAcc, the highest among the evaluated methods; the corresponding accuracy-difference intervals include zero (Appendix B.8, Table 10). For Qwen3-14B, only the label-free rule exceeds the logit baseline on both targets. Under the same protocol, the linear readout reaches 86.67% HLE BAcc for DeepSeek-Llama-8B and the annealed ensemble 90.92% MATH BAcc for Qwen3-32B.

Table 3: Transfer with 400 source labels: BAcc (%). MATH uses 100 held-out examples. Fitted rules are frozen. Best and second-best scores in each row are in bold and underlined, respectively.

Model	Dataset	Logits	Single QK	CF-QK	SA-QK	QK (L2)
Qwen3-8B	GSM8K	58.08	65.11	65.29	**69.34**	$\underline{66.39}$
	SVAMP	65.94	47.32	$\underline{67.39}$	**72.75**	65.53
Qwen3-14B	GSM8K	66.14	$\underline{66.83}$	**68.12**	54.94	65.27
	SVAMP	$\underline{64.14}$	62.48	**64.83**	59.73	59.73
DeepSeek-Llama-8B	MATH (100)	66.40	$\underline{82.39}$	67.21	81.98	**85.31**
	HLE	68.26	$\underline{84.58}$	74.98	74.31	**86.67**
Qwen3-32B	MATH (100)	70.25	87.48	84.76	**90.92**	$\underline{90.11}$
	HLE	74.81	76.99	$\underline{77.91}$	**78.91**	77.43

Results on all 500 MATH examples and two-pass controls appear in Appendices B and D.

Verification cost. CF-QK-R1P and SA-QK each score an already-generated candidate in one verifier pass; the counterbalanced variant requires two. SA search operates entirely on cached head scores. On a single H100 NVL, Qwen3-8B scores approximately 5.5 candidate solutions per second, and SA calibration takes approximately 105 seconds. These measurements exclude candidate generation and depend on sequence length and batching (Appendix B.9).

7 ANALYSIS

We examine which components support QK verification, how it depends on calibration and read positions, and whether the scores respond to solution content. The primary ablations evaluate one-pass CF-QK on the same examples as Table 1; analyses of the two-pass variant and calibrated readouts are identified separately.

7.1 AGGREGATION AND REFERENCE BATCHES

Agreement and responsiveness. The full one-pass weighting rule achieves the highest BAcc in five of eight model–dataset settings (Appendix B.4, Table 6). Removing responsiveness reduces MATH BAcc from 84.22% to 51.49% for Qwen3-14B, from 83.11% to 50.89% for Qwen3-32B, and from 75.41% to 50.00% for DeepSeek-Llama-8B. Responsiveness alone preserves much of the signal, and the full rule improves on it in six settings. Uniform weighting reaches 80.00% on Qwen3-8B MATH, compared with 79.36% for the full rule. The component comparison thus identifies responsiveness as particularly important for the larger Qwen models and DeepSeek-Llama-8B.

Independent reference and reference size. On the same evaluation examples as Table 1, freezing CF-QK weights from 50 unlabeled source examples gives 77.10%, 83.84%, and 82.63% MATH BAcc for Qwen3-8B, 14B, and 32B, within 2.26 points of their transductive scores. DeepSeek-Llama-8B reaches 68.43%. HLE source-frozen BAcc is 75.20–78.33% across all four models. Across ten 32-example source draws, HLE SD is 0.12–1.03 points; MATH SD is 0.28–1.54 on Qwen3 and 5.21 on DeepSeek-Llama-8B (Appendix B.6, Table 8). These SDs measure reference selection on fixed evaluation examples, separately from Table 1's bootstrap uncertainty. Appendix B.5 reports the separate 100-example MATH reference experiment.

7.2 CALIBRATION, READ POSITIONS, AND HEAD CONTRIBUTIONS

Calibration budget and complementary heads. Beyond the repeated-calibration comparison in Section 6.2, the Qwen3-8B budget sweep evaluates L2 on a fixed 100-example MATH holdout. Mean performance and candidate-pool overlap increase with more labels (Appendix C.6). Appendix C.4 shows head-level error patterns; Appendix C.7 decomposes an L2 score into all 128 head contributions and its intercept. Paired gains and split-level wins for the main repeated experiment appear in Appendix C.5.

Read positions, layers, and option mappings. The separate two-pass sweep compares the nine-view rule with individual anchors and layer pools (Appendix B.7, Table 9). The canonical colon-to-end-of-line view improves BAcc in three of four settings. Restricting this variant to the final quarter of layers gives 73.77% and 84.51% MATH BAcc for Qwen3-8B and 14B, versus 67.65% and 81.02% with the fixed final-half pool. The main one-pass method retains its label-free defaults throughout. In the separate DeepSeek-Llama-8B 500-example evaluation, adding the reversed option mapping raises CF-QK BAcc from 74.67% to 84.08% on MATH and from 74.98% to 84.41% on HLE (Appendices B.3 and D.1).

Complementary selection controls. MCQA experiments vary deliberation, option order, and Q/K extraction. On Qwen3-8B MMLU-Pro, native normalization raises direct QK accuracy from 43.2% to 45.5%, whereas CoT readouts reach 54.6% and 54.7% (Table 18). After CoT, QK and output accuracy differ by 0.4–0.9 points across the eight MMLU-Pro models (Appendix E.2); Appendix E.3 reports the HLE-MC and permutation controls. These option-selection experiments use task-specific readouts.

7.3 MATCHED CHANGES TO REASONING AND ANSWERS

We first construct 40 intact/corrupted pairs per model and dataset, changing one intermediate equality on MATH while preserving the final answer, or changing only the final option on HLE. Median character identity exceeds 99.94% (Appendix C.1). One-pass CF-QK prefers the intact MATH trace in 70% of Qwen3-8B pairs and 85% of Qwen3-14B pairs (one-sided sign tests $p = 0.0083$ and 4.2×10^{-6}), and the intact HLE trace in 100% and 95%. Verdict logits respond as well, with higher

MATH preference rates. At its zero threshold, however, one-pass CF-QK classifies the intact and corrupted MATH traces with only 51.25% (8B) and 58.75% (14B) accuracy; Table 11 gives the full comparison.

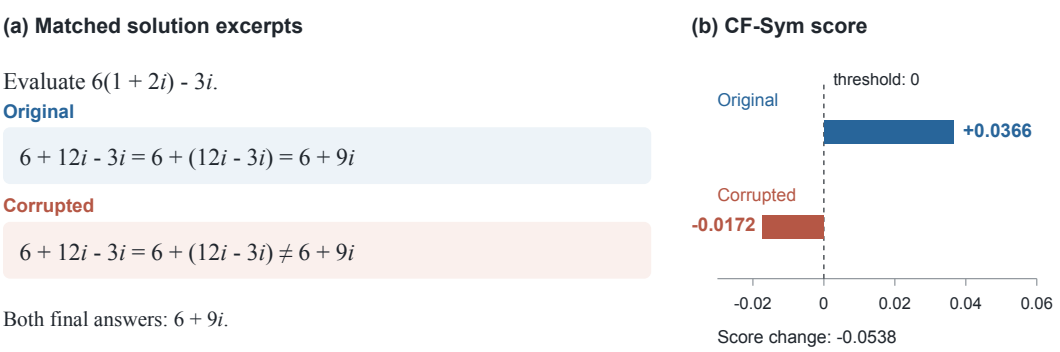


Figure 1: A matched Qwen3-8B MATH example. Changing one equality to an inequality preserves the final answer but moves the two-pass CF-Sym score from 0.0366 to -0.0172, across the zero threshold. All other solution text is unchanged.

A separate synthetic test holds the correct final answer fixed across 24 erroneous edits and 16 valid rewrites. On both models, CF-QK accepts every original and rejects all erroneous edits, while accepting both members of every valid pair. These results hold for full solutions and after removing the final-answer declaration; verdict logits achieve the same counts. Answer-only inputs are identical within each pair and yield no separation of erroneous pairs. Appendix C.2 describes the three error types, controls, and unlabeled transductive reference batch, with pair-level results in Table 12.

On the original MATH holdout, CF-QK also exceeds a cross-fitted lexical classifier using TF–IDF, overlap, coverage, and length features. The paired-bootstrap interval excludes zero for Qwen3-8B but not Qwen3-14B (Appendix C.1).

7.4 Causal influence on verdict scores

We test whether the heads selected by SA-QK affect the model's verdict. For Qwen3-8B, we patch queries, keys, or both between 80 intact/corrupted pairs (40 MATH and 40 HLE), averaging across the two verdict mappings. The outcome is the semantic next-token verdict-logit margin $m = z_{\text{correct}} - z_{\text{error}}$, rather than the QK score itself. With F and C denoting intact and corrupted runs, the bidirectional effect is

$$E = \tfrac{1}{2}\left[(m_{C\leftarrow F} - m_C) + (m_F - m_{F\leftarrow C})\right]. \tag{8}$$

Positive values indicate that intact activations increase the correct-verdict margin relative to corrupted activations, averaged over both intervention directions. This analysis uses the 11-head ensemble from the original 400-label calibration. We compare it with 11 random heads from the final half of the model, at most one random head per layer, pairing each query with its grouped-query key.

Joint Q+K patching gives a mean effect of 0.34 for selected heads and 0.04 for random heads. Their paired difference is 0.30 (95% CI $[0.24, 0.36]$). Key-only and joint interventions show positive selected-minus-random effects on both datasets; query-only effects differ in sign (Figure 2). Under these perturbations, the selected components influence the model's verdict, including when the final MATH answer is unchanged. Appendix C.3 reports the full controls and an alternative head-selection protocol.

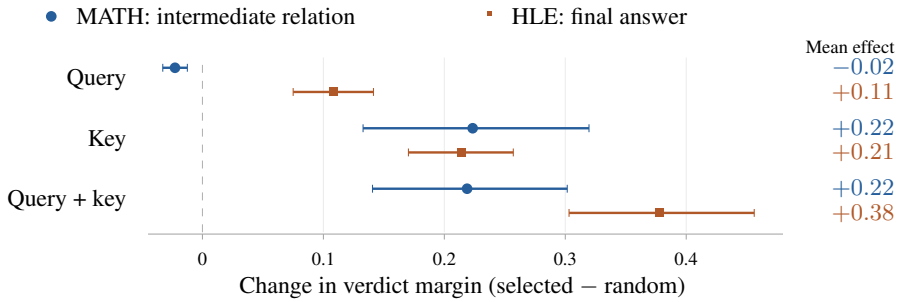


Figure 2: Causal effects of query and key patching in Qwen3-8B. Points show mean selected-minus-random bidirectional effects on the semantic verdict-logit margin; intervals are 95% paired-bootstrap intervals (40 pairs per dataset). Key and joint interventions have positive effects in both settings.

8 Conclusion

QK features support verification through unlabeled aggregation and lightweight supervised readouts, without updating the language model. We compare several ways to combine heads: CF-QK weights them using unlabeled agreement and responsiveness, while calibrated methods select heads or fit a regularized linear classifier. Across repeated 50-label calibration, L2-QK has higher mean BAcc than source-calibrated verdict logits for all four models on both MATH and HLE. The zero-label comparison and repeated fits characterize verification under distinct supervision budgets. Controlled perturbations and activation patching show that QK scores respond to changes in solution content and that selected heads influence the model's verdict.

Limitations

Our method requires white-box access to query and key projections and does not apply to API-only models. Labels refer to final-answer correctness: MATH labels come from an answer-equivalence judge, and HLE-1/4 is an author-constructed four-choice adaptation; step-level validity is examined only through controlled perturbations. CF-QK weights heads from an unlabeled batch of candidate solutions, and the calibrated readouts depend on the source split; Appendix B reports the resulting variation. Some dependence on prompt format and tokenization likely remains despite explicit delimiters and fixed read positions. The experiments cover mathematical reasoning and a text-only HLE adaptation with 8B–32B models, and the causal analysis covers one model.

AI use statement

In this work, we used generative AI tools to draft and edit parts of the manuscript, to identify related literature, to format references, and to propose the title. We have not used generative AI tools for any task with required disclosure under the ICLR 2027 AI Policy for Authors; the language models evaluated in the experiments are objects of study, and their roles are specified in Appendix A. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics and societal impact statement

This work aims to advance the field of machine learning, particularly reasoning and verification. There are many potential societal consequences of advances in this field; we do not identify any specific to the methods presented here that require separate discussion.

Reproducibility statement

Appendix A specifies prompts, extraction, aggregation, and data splits. Appendices B–E provide detailed results and selection protocols. We specify the number of calibration labels, the use of

unlabeled target examples, and the extraction convention for each experiment. Reported scoring costs exclude candidate generation.

REFERENCES

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14113–14130. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.781. URL https://aclanthology.org/2024.emnlp-main.781/.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision, 2022. URL https://arxiv.org/abs/2212.03827.

Shijie Chen, Bernal Jiménez Gutiérrez, and Yu Su. Attention in large language models yields efficient zero-shot re-rankers. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.02642.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. arXiv:2407.07071.

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

Sebastian Farquhar, Vikrant Varma, Zachary Kenton, Johannes Gasteiger, Vladimir Mikulik, and Rohin Shah. Challenges with unsupervised LLM knowledge discovery. arXiv:2312.10029, 2023. URL https://arxiv.org/abs/2312.10029.

Vipul Gupta, David Pantoja, Candace Ross, Adina Williams, and Megan Ung. Changing answer order can decrease mmlu accuracy, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. arXiv:2103.03874, 2021. URL https://arxiv.org/abs/2103.03874.

Guan Zhe Hong, Nishanth Dikkala, Enming Luo, Cyrus Rashtchian, Xin Wang, and Rina Panigrahy. A Implies B: Circuit analysis in LLMs for propositional logical reasoning. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2411.04105.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.01798.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-N selection for large language models via self-certainty. arXiv:2502.18581, 2025. URL https://arxiv.org/abs/2502.18581.

S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983. doi: 10.1126/science.220.4598.671.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. arXiv:2205.14334, 2022. URL https://arxiv.org/abs/2205.14334.

Valentin Noël. Geometry of reason: Spectral signatures of valid mathematical reasoning. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2601.00791.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.02707.

Sophie Ostmeier, Brian Axelrod, Maya Varma, Asad Aali, Yabin Zhang, Magdalini Paschali, Sanmi Koyejo, Curtis Langlotz, and Akshay Chaudhari. Gradient-stable attention heads signal LLM correctness. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, 2026. arXiv:2602.13699.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, 2021. URL https://aclanthology.org/2021.naacl-main.168/.

Long Phan et al. Humanity's last exam. arXiv:2501.14249, 2025. URL https://arxiv.org/abs/2501.14249.

Qwen Team. Qwen2.5-Math-PRM-7B. Model card, 2025. URL https://huggingface.co/Qwen/Qwen2.5-Math-PRM-7B.

Arnab Sen Sharma, Giordano Rogers, Natalie Shapira, and David Bau. LLMs process lists with general filter heads. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.26784.

Eduard Tulchinskii, Laida Kushnareva, Kristian Kuznetsov, Anastasia Voznyuk, Andrei Andriiainen, Irina Piontkovskaya, Evgeny Burnaev, and Serguei Barannikov. Listening to the wise few: Select-and-copy attention heads for multiple-choice QA. arXiv:2410.02343, 2024. URL https://arxiv.org/abs/2410.02343.

Eduard Tulchinskii, Anastasia Voznyuk, Laida Kushnareva, Andrei Andriiainen, Irina Piontkovskaya, Evgeny Burnaev, and Serguei Barannikov. Quantifying logical consistency in transformers via query-key alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, 2025. arXiv:2502.17017.

Gladys Tyen, Hassan Mansoor, Victor Cărbune, Peter Chen, and Tony Mak. LLMs cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. arXiv:2311.08516.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.15098.

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. In *Conference on Language Modeling (COLM)*, 2025a. arXiv:2504.05419.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. Query-focused retrieval heads improve long-context reasoning and re-ranking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025b. arXiv:2506.09944.

Zheng Zhao, Yeskendir Koishekenov, Xianjun Yang, Naila Murray, and Nicola Cancedda. Verifying chain-of-thought reasoning via its computational graph. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.09312.

Table 4: Evaluation settings. Verification predicts candidate correctness; multiple-choice selection predicts an answer. Dataset sizes refer to evaluation unless otherwise stated. HLE-1/4 and HLE-MC are author-constructed four-choice adaptations of HLE.

Task	Evaluation data	Calibration and readout
Main verification	MATH: 450 per model; HLE: 2,156	CF: unlabeled target batch, zero threshold. QK calibration: 50 MATH labels; scalar baselines: the same 50-label source budget. Scores and thresholds frozen at transfer.
Original calibration and transfer	MATH: 100 per model; HLE: 2,156; GSM8K: 1,319; SVAMP: 300	400 MATH calibration labels; fitted rules frozen at transfer.
Cross-family verification	DeepSeek-Llama-8B: MATH 500; HLE 2,156	Same CF rule; no correctness labels.
Multiple-choice selection	MMLU-Pro: 2,000; HLE-MC: 1,600 per split	500 calibration questions; five splits; before/after CoT.

A Implementation details

A.1 Verification prompt and activation extraction

The verifier receives the problem and candidate solution, but not the reference answer:

```
You are verifying whether a candidate reasoning chain is
correct.
Problem: [problem]
Candidate solution: [solution]
Choose the correct verdict:
A. The candidate solution is correct.
B. The candidate solution contains an error.
Verdict:
```

The second mapping interchanges the two option sentences. Query anchors are the first and last characters of `Verdict` and its colon; key anchors are the final content character, terminal period, and following newline of each option. Character offsets are mapped to tokenizer positions. The nine nominal views are deduplicated when their complete measurements coincide.

The verification implementation captures *query/key projection outputs*, before native Q/K normalization where present and before rotary position encoding. It reshapes these outputs into heads and takes their raw dot products, with neither softmax nor $1/\sqrt{d}$ scaling. For grouped-query attention, query head h uses key head $\lfloor h/g \rfloor$, where g is the number of queries per key. Inputs exceeding 4,096 tokens are shortened by removing tokens from the middle of the candidate solution; the problem and verdict options are preserved. The MATH/HLE launch configuration defaults to FP16; the GSM8K/SVAMP and cross-family configurations default to BF16. Extracted scores are stored in FP32.

A.2 Responsive query–key aggregation

For a mapping set M, let $\mathcal{V}_M(x)$ contain the unique mapping–view pairs for example x. Given the semantic margins in Eq. (2), each head casts the mean preference vote

$$u_h(x) = \frac{1}{|\mathcal{V}_M(x)|} \sum_{(m,v)\in\mathcal{V}_M(x)} \operatorname{sign}\big(d_{h,m,v}(x)\big). \tag{9}$$

Exact raw-margin ties contribute zero. With $\epsilon = 10^{-8}$ and unlabeled reference batch $\mathcal{B}$, define

$$r_h(x) = \mathbf{1}[|u_h(x)| > \epsilon], \qquad c_h = |\mathcal{B}|^{-1} \sum_{x\in\mathcal{B}} r_h(x), \tag{10}$$

$$a_h = |\mathcal{B}|^{-1} \sum_{x\in\mathcal{B}} |u_h(x)|, \qquad p_h = \frac{\sum_{x\in\mathcal{B}} \mathbf{1}[u_h(x) > 0] r_h(x)}{\sum_{x\in\mathcal{B}} r_h(x)}, \tag{11}$$

$$w_h = \mathbf{1}[\ell(h) \geq \lfloor L/2 \rfloor] 4 p_h (1 - p_h) c_h a_h. \tag{12}$$

Coverage c_h is the fraction of non-tied *mean head votes*; agreement a_h includes tied examples. If the denominator defining p_h vanishes, set $p_h = 1/2$, leaving its weight zero through coverage. If $\sum_h w_h \leq \epsilon$, use uniform weights over eligible heads. The aggregate is $S_M(x) = \sum_h w_h u_h(x) / \sum_h w_h$. Predict correct only for $S_M(x) > 0$; zero is incorrect. The main CF-QK experiments did not require this fallback.

CF-QK-R1P computes S_1 from the mapping A=correct. CF-QK-RSym computes S_{12} by pooling both mappings and recomputing their batch weights, then returns

$$S_{\mathrm{RSym}} = \tfrac{1}{2}(S_1 + S_{12}). \tag{13}$$

This two-pass average still depends on which mapping is presented first. The position and layer ablations evaluate CF-QK-RSym; the main one-pass comparison evaluates CF-QK-R1P.

The stability-only aggregation baseline operates per example: it converts scores to normalized Borda ranks and weights all heads by nonnegative mean pairwise Kendall agreement across views. Its two-pass version additionally weights semantic agreement between mappings. It aggregates median Borda scores across views and uses uniform head weights when all stability weights vanish. A head that always prefers one verdict can receive a high stability weight; the responsiveness factor in CF-QK-R1P and CF-QK-RSym suppresses such constant preferences.

A.3 CF-QK DESIGN RATIONALE

The CF-QK defaults were chosen without correctness labels. This applies to the layer cutoff, query/key anchors, responsiveness formula, zero threshold, and coefficient of the two-pass blend. No reported evaluation labels were consulted to select these defaults. Labeled search and validation are confined to the separately identified calibrated readouts.

Layer pool and read anchors. Hong et al. (2025) identify distinct attention-head functions within propositional-reasoning circuits, while Biran et al. (2024) show that the second step of a multi-hop query is resolved in later layers. These findings motivate concentrating the CF-QK pool in later layers; the final-half cutoff is our fixed implementation of this choice. The read anchors follow the prompt structure: the verdict boundary follows the complete candidate solution, and option-end keys follow the corresponding correctness statement. Nearby token anchors provide multiple views of these same semantic boundaries.

Responsiveness and the two-pass blend. The factor $4p_h(1-p_h)$ is the normalized variance of a binary head preference: it vanishes for a constant verdict and reaches one at $p_h = 1/2$. Coverage and agreement further weight the available, consistent votes. These statistics are computed from unlabeled activations. The two-pass rule combines the first-mapping score with a score pooling both mappings using the fixed coefficient $1/2$. It preserves a first-mapping component while incorporating the reversed options; it is a separate variant from the primary one-pass rule. Computing reference weights on the unlabeled batches specified above does not fit these design choices to correctness labels.

A.4 SA-QK ENSEMBLE SEARCH

SA-QK uses the canonical colon-to-end-of-line margin in the A=correct mapping. It divides each head's margin by its search-set IQR *plus* 10^{-8}, without centering. The candidate pool contains the 64 highest individual search-AUROC heads and 64 complementary heads from across the model. Complementary heads minimize maximum absolute correlation with the selected pool, preferring nonconstant heads with AUROC at least 0.5; individual AUROC and head index break ties. If too few exceed chance, further nonconstant heads are admitted. Correlation uses centered, unit-norm columns, but the ensemble score remains the equal-weight mean of IQR-scaled margins.

For each $K \in \{3, 5, 7, 9, 11\}$, 20 restarts make 10,000 swaps within this pool. Restart zero starts from the top-K heads, restart one from a diverse subset, and the remainder from random subsets. Among 200 random swap comparisons, let D be the median magnitude of worsening AUROC changes, or 10^{-3} if none occur. The schedule is

$$T_0 = \max\{D/[-\log(0.8)], 10^{-8}\}, \qquad T_t = T_0(10^{-3})^{t/9999}, \quad t = 0, \ldots, 9999. \tag{14}$$

Non-worsening moves are accepted; other moves have probability $\exp\{\Delta/\max(T_t, 10^{-8})\}$. Each restart retains its best search-AUROC subset. Validation AUROC selects the restart and size, with search AUROC and head index breaking ties. The threshold is fitted separately to maximize balanced accuracy over score midpoints and the two extreme decisions; ties favor ordinary accuracy, then proximity to zero. All subset search uses cached scores.

A.5 Data splits and transfer

Fixed 50-label calibration split. The 50 calibration examples are allocated to 30 search, 10 validation, and 10 threshold examples, using seed $42 + 100 + r$ for role index r. The QK scorers fitted with this allocation are held fixed in the comparison. Scalar baselines retain their scores and recalibrate their thresholds with the same 50-label source budget. The remaining 450 MATH examples form a separate evaluation set and are excluded from calibration. HLE uses all 2,156 examples without target-label calibration.

Repeated 50-label calibration. Table 2 repeats the 50-label allocation over ten stratified splits and three search seeds per split for each model. The three seed results are averaged within each split before computing the reported mean and SD over ten splits. Each split uses its remaining 450 MATH examples for evaluation; HLE remains fixed. Appendix C.5 reports paired comparisons and distinguishes calibration variation from conditional evaluation uncertainty.

Original 400-label experiment. MATH-500, drawn from MATH (Hendrycks et al., 2021), is divided into 240 search, 80 validation, 80 threshold and 100 evaluation examples. The outer stratified split uses seed 42; the two inner stratified splits use seeds 43 and 44. The annealing master seed is 42. Qwen3-8B and Qwen3-14B both select $K = 11$; thresholds are 0.2407638394 for 8B and -0.5155734037 for 14B. Heads, scales and thresholds transfer unchanged to target datasets, so SA-QK uses 400 source labels and zero target labels. All Qwen verifiers in this experiment use the same 100 evaluation examples within each model; stratification means that the two models need not share identical evaluation indices.

The selected zero-based $(\text{layer}, \text{head})$ pairs for Qwen3-8B are $(9, 11)$, $(9, 24)$, $(13, 8)$, $(23, 22)$, $(24, 21)$, $(24, 30)$, $(26, 1)$, $(26, 6)$, $(26, 31)$, $(33, 14)$, and $(34, 14)$. For Qwen3-14B they are $(1, 2)$, $(5, 2)$, $(13, 2)$, $(13, 5)$, $(21, 16)$, $(21, 36)$, $(22, 33)$, $(27, 9)$, $(36, 25)$, $(36, 29)$, and $(38, 22)$.

In the 400-label comparison, CF-QK-R1P and CF-QK-RSym compute weights from all 500 unlabeled MATH examples, including the scored examples, while reporting on the 100-example holdout. HLE-1/4 uses all 2,156 examples for both unlabeled weighting and evaluation. The initial DeepSeek-Llama-8B CF-QK evaluation uses all 500 MATH examples and all 2,156 HLE examples. Qwen3-8B transfer uses 1,319 GSM8K and 300 SVAMP examples. Scoring uses these unlabeled reference batches; CF-QK uses correctness labels only for evaluation.

A.6 Correctness labels and baseline scoring

MATH correctness is determined by a Qwen3-32B answer-equivalence judge; HLE-1/4 uses exact option-letter matching. The deterministic four-option construction from HLE (Phan et al., 2025) uses seed 42. Labels refer to final-answer correctness. Direct verdict logits are read from the same passes as QK; the two-pass logit baseline averages semantic A/B margins. Indirect confidence uses the zero-shot True-minus-False logit margin; generated-solution likelihood is the mean conditional log-probability of the frozen reasoning chain. In the original experiment, calibrated scalar baselines fit their threshold on the 400 MATH calibration examples and transfer it unchanged. In the main experiment, the scalar baselines use the same 50-label source budget as the QK readouts. The external PRM comparison has a separate 2,048-token limit and at most 32 uniformly retained paragraph steps. Accuracy-difference intervals use 10,000 paired percentile-bootstrap resamples.

For the non-judge entries with bootstrap uncertainty in Tables 1 and 5, we compute standard deviations over 10,000 paired bootstrap resamples stratified by correctness. Each resample preserves the number of correct and incorrect examples and uses the same sampled indices across methods. We use the sample standard deviation (`ddof=1`). Calibration parameters, decision thresholds, and CF-QK reference weights remain fixed. These standard deviations quantify conditional evaluation uncertainty.

The separate experiment in Appendix C.6 measures variation across source calibration splits on a fixed MATH evaluation set.

Open-generation judge. For Qwen judges, Table 5 reports $(L + U)/2$, the midpoint of the lower and upper BAcc bounds for unparsed outputs. This retains all evaluation examples without assigning a verdict to each unparsed response. No SD is reported for the midpoint because the endpoint SDs do not determine it without their covariance. Parsed counts for Qwen3-8B, Qwen3-14B, and Qwen3-32B are 436, 443, and 431 of 450 MATH examples, and 2,091, 2,129, and 2,039 of 2,156 HLE examples, respectively.

DeepSeek-Llama-8B produces valid verdicts for 446 of 450 MATH examples and 2,130 of 2,156 HLE examples. Unparsed verdicts count as incorrect predictions, and all evaluation examples remain in the metric. Its BAcc is $54.82 \pm 1.22\%$ on MATH and $68.32 \pm 1.02\%$ on HLE, with the reported SDs included in Table 5.

A.7 Verification datasets and candidate solutions

The verification datasets and candidate solutions are fixed before score comparison. MATH uses the test split of `HuggingFaceH4/MATH-500`; HLE-1/4 uses a deterministic four-choice construction from `krammnic/hle-multichoice`. With seed 42, existing multiple-choice items retain the correct answer and three alternatives, whereas open-ended items use the three distractors supplied by that dataset. Choices are shuffled and relabeled A–D; entries with missing or unusable options are omitted. The resulting 2,156 questions and reference choices are identical across the verification models. The MATH question/reference pairs likewise match across models, while each model supplies its own candidate solutions. HLE labels use option-letter equality; GSM8K and SVAMP labels use deterministic scalar numeric equivalence. The 2,100-question HLE-MC collection in Appendix E is a separate evaluation set.

The stored prediction files retain the problem, reference answer, candidate solution, extracted final answer, and correctness label. The primary MATH files additionally retain the raw answer-equivalence judge output. These fixed candidates are shared by all scoring methods within each model/dataset comparison, so differences between scorers do not reflect resampling of solutions.

B Additional verification results

We examine how the verification results depend on model family, head aggregation, read positions, and task distribution. All experiments follow the extraction, data-splitting, and labeling protocols in Appendix A. Accuracy, balanced accuracy (BAcc), and AUROC are reported in percent. The ensemble and transfer results use the original 400-label protocol; the independent-reference evaluation uses no calibration labels. The one-pass weighting ablation uses the main 450-example MATH evaluation set. The original ensemble and transfer comparisons use a shared 100-example MATH holdout within each Qwen model; those indices differ between Qwen3-8B and Qwen3-14B. The HLE evaluation contains 2,156 four-option examples with exact option-letter correctness labels.

B.1 Verification with a fixed calibration split

Table 5 reports the full comparison on the original 50-label calibration split, with 450 fixed MATH evaluation examples and 2,156 HLE examples per model. Each calibrated rule is fitted once; its bootstrap SD holds that fit fixed. These results are reported separately from the repeated-calibration means and split SDs in Table 2. The zero-label logits and CF-QK entries also appear in Table 1. Appendix C.5 reproduces the original Qwen3-32B fit and examines its decision threshold.

Table 5: Verification with a fixed 50-label calibration split. All calibrated methods use 50 source labels. BAcc (%), with SD where available (see note). MATH: 450 examples; HLE: 2,156. In each column, the best and second-best estimates within the zero-label and calibrated blocks are in bold and underlined, respectively.

		Qwen3-8B		Qwen3-14B	
Method	Labels	MATH (450)	HLE	MATH (450)	HLE
LLM-as-judge[a]	0	61.72	66.98	63.54	68.32
Verdict logits	0	<u>65.74</u> ± 1.71	75.61 ± 2.27	<u>75.50</u> ± 1.97	**80.48** ± 1.96
Indirect logits	0	65.60 ± 1.74	<u>77.56</u> ± 2.63	68.80 ± 1.89	<u>77.66</u> ± 2.02
CF-QK-R1P	0	**79.36** ± 1.93	**80.84** ± 2.60	**84.22** ± 1.79	74.79 ± 2.21
Verdict logits (cal.)	50	74.91 ± 1.91	70.92 ± 3.08	77.07 ± 1.63	54.13 ± 1.43
Indirect logits (cal.)	50	74.77 ± 1.93	72.54 ± 3.09	71.43 ± 2.13	60.70 ± 2.09
Log-probability (cal.)	50	54.18 ± 2.34	48.85 ± 0.16	61.53 ± 2.22	49.25 ± 0.65
Single QK head	50	78.33 ± 1.86	72.12 ± 2.85	83.12 ± 1.81	74.35 ± 2.25
Top-K QK mean	50	**83.67** ± 1.77	<u>75.79</u> ± 2.78	84.09 ± 1.75	73.59 ± 2.29
Greedy subset search	50	69.08 ± 2.05	69.82 ± 3.04	77.24 ± 1.61	78.47 ± 2.15
Random subset search	50	75.27 ± 2.02	67.68 ± 3.03	82.30 ± 1.68	**79.76** ± 2.10
SA-QK	50	66.89 ± 1.88	65.60 ± 2.42	<u>84.86</u> ± 1.50	<u>79.65</u> ± 2.14
Linear QK (L2)	50	<u>83.65</u> ± 1.76	**79.73** ± 2.71	**86.48** ± 1.53	78.95 ± 2.17

		Qwen3-32B		DeepSeek-Llama-8B	
Method	Labels	MATH (450)	HLE	MATH (450)	HLE
LLM-as-judge[a]	0	56.54	60.82	54.82 ± 1.22	<u>68.32</u> ± 1.02
Verdict logits	0	<u>65.26</u> ± 1.81	<u>74.81</u> ± 1.63	<u>63.55</u> ± 1.64	68.26 ± 1.44
Indirect logits	0	60.00 ± 1.56	71.67 ± 1.71	58.57 ± 1.30	59.90 ± 2.15
CF-QK-R1P	0	**83.11** ± 1.76	**77.91** ± 1.87	**75.41** ± 1.95	**74.98** ± 2.29
Verdict logits (cal.)	50	79.28 ± 1.97	**76.63** ± 2.16	79.87 ± 1.81	52.31 ± 1.05
Indirect logits (cal.)	50	72.03 ± 2.12	62.30 ± 2.03	65.91 ± 2.20	61.38 ± 2.33
Log-probability (cal.)	50	56.72 ± 2.30	48.26 ± 0.45	46.72 ± 2.02	36.64 ± 1.72
Single QK head	50	**82.76** ± 1.81	<u>75.02</u> ± 2.15	81.19 ± 1.86	74.39 ± 1.04
Top-K QK mean	50	**82.76** ± 1.81	<u>75.02</u> ± 2.15	83.19 ± 1.77	63.80 ± 2.09
Greedy subset search	50	<u>80.69</u> ± 1.93	52.62 ± 2.30	79.78 ± 1.80	78.46 ± 0.93
Random subset search	50	73.30 ± 2.11	67.29 ± 2.29	75.60 ± 1.86	71.71 ± 0.71
SA-QK	50	74.84 ± 2.09	62.74 ± 2.27	<u>83.44</u> ± 1.74	**85.80** ± 1.00
Linear QK (L2)	50	73.43 ± 1.99	68.19 ± 1.27	**88.44** ± 1.53	<u>84.97</u> ± 1.50

[a]Qwen judge scores are midpoints of the lower and upper BAcc bounds for unparsed outputs; their SD cannot be recovered from the endpoint SDs without their covariance. DeepSeek judge scores include the reported SD, counting unparsed verdicts as incorrect. SDs for the other methods use 10,000 bootstrap resamples.

B.2 Ensemble performance

Table 15 compares verdict logits, calibration-free aggregation, and SA-QK, including the two-pass variants and DeepSeek-Llama-8B results. SA-QK uses 400 MATH labels for head selection and threshold fitting, while CF-QK estimates its responsiveness weights from all 500 unlabeled MATH examples. At transfer, SA-QK keeps the source heads, scales, and threshold fixed; CF-QK recomputes its unlabeled batch weights. Neither procedure uses target HLE labels.

The selected Qwen3-8B ensemble contains 11 heads and achieves 95.36% search AUROC, 93.40% validation AUROC, and 89.23% held-out AUROC. Its held-out accuracy improvement over matched one-pass verdict logits is 15.00 percentage points, with paired-bootstrap 95% CI $[2.00, 28.00]$. The 14B ensemble also improves MATH verification, reaching 95.28% AUROC, but its HLE accuracy and BAcc fall below the logit baseline.

Class imbalance. Balanced accuracy accounts for the class imbalance in HLE. On Qwen3-8B HLE, a MATH-calibrated likelihood threshold attains 93.78% accuracy but 49.11% BAcc, while CF-QK attains 84.18% accuracy and 80.84% BAcc. The likelihood baseline achieves higher accuracy but rarely identifies correct solutions at this threshold.

B.3 Transfer across model families

We evaluate the same calibration-free rule on DeepSeek-Llama-8B using one or two verdict mappings. The two-pass responsive score defined in Appendix A is

$$S_{\mathrm{RSym}}(x) = \tfrac{1}{2}\big(S_1(x) + S_{12}(x)\big). \tag{15}$$

Here S_1 uses the first mapping and S_{12} pools both mappings before computing responsiveness weights. This average still depends on which mapping is presented first. The matched two-pass logit baseline averages the semantic verdict margins.

In Table 15, the two-pass CF-QK variant improves BAcc over its matched logit baseline on both datasets. Its AUROC is lower on MATH and higher on HLE, showing that the benefit depends on the metric and task. DeepSeek-Llama-8B evaluation uses all 500 MATH examples; the Qwen evaluation uses a 100-example holdout.

B.4 One-pass weighting ablations

We ablate CF-QK-R1P on exactly the evaluation rows used in Table 1: 450 MATH examples and 2,156 HLE examples per model. MATH reference weights use all 500 unlabeled examples, including the evaluation rows. HLE uses the full unlabeled evaluation batch. Eligible heads, the nine query–key views, and the zero decision threshold are held fixed. Table 6 compares the full rule with uniform weighting, agreement-only weighting, responsiveness-only weighting, and removal of responsiveness.

The full-rule BAcc matches Table 1 in all eight settings. It has the highest BAcc in five settings; uniform weighting leads on Qwen3-8B MATH, and responsiveness alone leads on Qwen3-14B MATH and DeepSeek-Llama-8B HLE. Removing responsiveness reduces Qwen3-14B MATH BAcc from 84.22% to 51.49%, Qwen3-32B MATH from 83.11% to 50.89%, and DeepSeek-Llama-8B MATH from 75.41% to 50.00%.

B.5 Independent reference batch

Table 7 reports one-pass CF-QK-R1P using an independent source reference batch, evaluated on 100 MATH examples and 2,156 HLE examples.

B.6 Reference-batch dependence on fixed evaluation examples

We compare one-pass CF-QK reference configurations on the same 450 MATH evaluation examples and the same 2,156 HLE examples within each model. The transductive configuration uses the unlabeled batches of the main experiment. Source-frozen 50 computes weights from the 50 source examples without their labels and holds them fixed for evaluation. Source-32 uses ten draws of 32 source examples. The head pool and zero threshold remain fixed. Table 8 reports one point

Table 6: One-pass CF-QK-R1P weighting ablations: BAcc (%) on the same 450 MATH and 2,156 HLE evaluation rows as Table 1. MATH weights use the full 500-row unlabeled reference batch. Best and second-best values in each row are in bold and underlined, respectively.

Model	Dataset	Full	Uniform	Agreement	Responsive	No response
Qwen3-8B	MATH (450)	79.36	**80.00**	<u>79.75</u>	79.30	78.52
	HLE	**80.84**	78.99	<u>80.55</u>	78.32	80.45
Qwen3-14B	MATH (450)	<u>84.22</u>	59.91	52.19	**84.63**	51.49
	HLE	**74.79**	63.62	55.19	<u>74.22</u>	54.07
Qwen3-32B	MATH (450)	**83.11**	59.64	51.78	<u>78.96</u>	50.89
	HLE	**77.91**	70.40	69.04	<u>74.84</u>	66.86
DeepSeek-Llama-8B	MATH (450)	**75.41**	50.00	50.00	<u>74.68</u>	50.00
	HLE	<u>74.98</u>	49.71	49.88	**77.06**	49.90

Table 7: One-pass CF-QK-R1P with an independent source reference batch: BAcc (%) $\pm$ SD.

Model	MATH (100)	HLE
Qwen3-8B	74.05 ± 4.38	80.38 ± 2.84
Qwen3-14B	82.37 ± 3.93	75.63 ± 2.30
Qwen3-32B	84.76 ± 3.40	77.41 ± 2.08
DeepSeek-Llama-8B	68.99 ± 4.60	77.47 ± 2.23

estimate for each deterministic configuration and mean $\pm$ SD across reference draws for Source-32. These SDs measure reference-batch variation, not evaluation-bootstrap uncertainty. This matched 450-example comparison is separate from the 100-example MATH experiment in Table 7.

For Qwen3, source-frozen MATH BAcc differs from the transductive score by 0.38–2.26 points. HLE source-frozen BAcc lies between 75.20% and 78.33% across the four models. Source-32 HLE SDs range from 0.12 to 1.03 points. DeepSeek-Llama-8B MATH shows a larger reference effect: 75.41% transductive, 68.43% with the fixed source reference, and 71.08 ± 5.21% across Source-32 draws.

Table 8: One-pass CF-QK reference-batch comparison: BAcc (%) on fixed MATH (450) and HLE (2,156) evaluation examples. Source-32 reports mean $\pm$ SD over ten unlabeled source draws; the other configurations are deterministic point estimates.

Model	Dataset	Transductive	Source-frozen 50	Source-32
Qwen3-8B	MATH	79.36	77.10	77.82 ± 1.54
	HLE	80.84	78.33	79.21 ± 1.03
Qwen3-14B	MATH	84.22	83.84	84.13 ± 0.74
	HLE	74.79	75.20	74.77 ± 0.96
Qwen3-32B	MATH	83.11	82.63	82.81 ± 0.28
	HLE	77.91	77.68	77.78 ± 0.12
DeepSeek-Llama-8B	MATH	75.41	68.43	71.08 ± 5.21
	HLE	74.98	75.40	75.99 ± 1.00

B.7 Two-pass read-position and layer-pool ablations

We vary the query and key positions and the eligible layers, using the same candidate solutions and evaluation examples. This sweep uses the two-pass responsive score in Eq. (15). The independent-reference results are in Appendix B.5. The canonical single view pairs the colon following `Verdict` with the end-of-line boundary after each verdict option. The full anchor set and deduplication rule are given in Appendix A.

The first two rows of Table 9 compare all nine views with the canonical single view. The single view improves BAcc in three of four settings, but reduces Qwen3-14B MATH AUROC by 5.62 points. One query–key pair can therefore provide an informative score. Table 9 reports the full position and

layer-pool sweep. The main experiments use heads from the final half of the layers on every dataset; we do not select a separate layer pool for each test set.

Table 9: Read-position and layer-pool sensitivity, including the nine-view and canonical single-view comparison in the first two rows. Each cell is BAcc (%); all rows use the two-pass responsive blend. "All Q" and "all K" retain all three anchors of that type. Position sweeps retain the final-half layer pool; layer sweeps retain all nine views. EOL and terminal punctuation yield identical extracted QK measurements in all four settings and are shown in one row; identical views are deduplicated before aggregation.

	Qwen3-8B		Qwen3-14B	
Readout	MATH	HLE	MATH	HLE
All nine views	67.65	78.21	81.02	77.20
Colon × EOL	72.64	79.00	78.23	78.12
Colon × all K	67.99	79.69	80.37	77.94
All Q × EOL/punctuation	73.19	80.63	76.17	75.29
Start × all K	66.77	75.75	80.31	69.91
End × all K	76.95	77.93	82.45	75.52
All Q × content	72.87	71.88	82.05	75.32
All layers	70.24	75.34	81.02	72.18
Last 75%	68.79	74.70	81.02	73.18
Last 50%	67.65	78.21	81.02	77.20
Last 25%	73.77	79.74	84.51	77.23

B.8 Transfer to mathematical reasoning benchmarks

We evaluate Qwen3-8B on 1,319 GSM8K examples (Cobbe et al., 2021) and 300 SVAMP examples (Patel et al., 2021). All methods score the same candidate traces. SA-QK transfers the MATH head ensemble, normalization, and threshold without target labels. The indirect-logit and likelihood baselines with a MATH threshold likewise use 400 source labels; their zero-shot variants and CF-QK use none. Table 10 reports the source-label budget for each method.

SA-QK has the highest BAcc on both datasets and the highest GSM8K AUROC among these methods; on SVAMP it essentially matches direct logits. The calibration-free rule has one-pass AUROC of 63.27 on GSM8K and 55.35 on SVAMP. In these experiments, the MATH-selected ensemble transfers better than CF-QK.

Table 10: Qwen3-8B verification balanced accuracy (%) on GSM8K and SVAMP. "Labels" counts source MATH labels; all methods use zero target labels. PRM uses a 2,048-token budget.

		GSM8K ($n = 1{,}319$)	SVAMP ($n = 300$)
Method	Labels	BAcc	BAcc
Verdict logits, one pass	0	58.08	65.94
Verdict logits, two passes	0	60.83	66.63
Indirect logits	0	61.16	67.52
Indirect logits, MATH threshold	400	65.56	70.14
Likelihood, MATH threshold	400	62.11	71.88
PRM-min	0	55.28	59.03
CF-QK, one pass	0	65.29	67.39
CF-QK, two-pass blend	0	63.79	68.49
SA-QK, frozen MATH	400	69.34	72.75

B.9 Scoring cost

On a single H100 NVL, Qwen3-8B QK scoring processes approximately 5.5 already-generated traces per second: 500 traces take approximately 90.9 seconds (1.52 minutes). Single-head calibration takes approximately 80 seconds and SA-QK calibration approximately 105 seconds; CF-QK has no labeled calibration stage. These measurements exclude solution generation. Annealing operates on cached

scores, while the two-pass readout requires two verifier passes. Throughput depends on batching and sequence length.

C Analysis of verification signals

We examine whether QK scores respond to reasoning changes beyond lexical overlap, and whether the corresponding attention components influence the verifier's verdict. Matched counterfactuals test sensitivity to localized changes; activation patching tests their effect on downstream verdict logits.

C.1 Matched counterfactuals

For each of Qwen3-8B and Qwen3-14B, we construct 40 MATH counterfactual pairs from correct solutions by changing exactly one final intermediate equality to an inequality. The problem, remaining reasoning, and final answer are unchanged. For 40 HLE pairs, only the final answer option is replaced; the preceding reasoning is unchanged. Median character identity exceeds 99.94%. Table 11 measures preference for the intact trace. The MATH intervention tests intermediate consistency while holding the final answer fixed, complementing the answer-correctness evaluation.

Table 11: Preference for intact counterfactual traces (%); 40 pairs per model–dataset cell. Ties count as half a preference.

	Qwen3-8B		Qwen3-14B	
Method	MATH	HLE	MATH	HLE
Verdict logits, one pass	85.00	100.00	93.75	95.00
Verdict logits, two passes	87.50	97.50	90.00	97.50
CF-QK, one pass	70.00	100.00	85.00	95.00
CF-QK, two-pass blend	85.00	100.00	87.50	97.50

One-pass CF-QK prefers the intact MATH trace above chance, with one-sided sign-test $p = 0.0083$ for 8B and $p = 4.2 \times 10^{-6}$ for 14B. Verdict logits achieve higher MATH pairwise preference rates, indicating that both internal signals respond to the perturbation. At the fixed zero threshold, one-pass CF-QK classifies the intact/corrupted MATH traces with 51.25% accuracy for 8B and 58.75% for 14B.

We compare the ordinary verification scores against a cross-fitted supervised lexical classifier using word/character TF–IDF similarity, token overlap, coverage, and length features. CF-QK has higher point-estimate AUROC in all four settings, although three paired-bootstrap intervals include zero. The clearest separation is on Qwen3-8B MATH, where the AUROC advantage is 13.46 points. The paired-bootstrap 95% intervals for the AUROC difference are $[3.29, 23.96]$ and $[-0.14, 18.06]$ points on MATH for 8B and 14B, respectively, and $[-3.78, 9.13]$ and $[-3.48, 5.70]$ on HLE. These intervals use 10,000 paired bootstrap resamples.

C.2 Reasoning perturbations with a fixed final answer

We test whether the verifier distinguishes an invalid reasoning step from a valid rewrite when the problem and final answer are unchanged. Unlike the equality-edit experiment in Appendix C.1, this test includes both erroneous edits and validity-preserving controls, and evaluates each pair with and without its reasoning.

Paired examples. We construct 40 synthetic problems, each with an original valid solution and an edited counterpart sharing the same correct final answer. In 24 pairs, the edit introduces an incorrect arithmetic equality, an invalid algebraic transformation, or a division that discards a possible zero root (eight pairs of each type). The remaining 16 pairs preserve validity: eight use a correct rewrite with inequality notation, and eight change only the wording. The examples are generated from templates reviewed by ourselves.

Each solution is presented in three forms: the full solution; the reasoning with the explicit final-answer declaration removed; and the final answer alone. The answer-only inputs are identical within each pair. Removing the final-answer declaration leaves the intermediate reasoning intact, including any numerical result stated there. This gives 240 inputs per model: 40 pairs, two solutions per pair, and three representations.

Scoring and evaluation. We use frozen Qwen3-8B and Qwen3-14B, comparing CF-QK with direct verdict logits. No supervised readout is fitted on these examples. CF-QK weights are computed without labels from the 80 full-solution inputs, including both members of each pair, and then held fixed across all three representations. This is an unlabeled transductive reference batch. Decisions use the prespecified zero threshold.

For an erroneous edit, a pair counts as successful only if the verifier accepts the original solution and rejects the edited one. For a validity-preserving edit, success requires accepting both. Table 12 reports these pair-level counts; they differ from the balanced-accuracy metric used in the main verification tables.

Table 12: Fixed-answer reasoning test. Entries count successful pairs out of 24 erroneous edits or 16 valid edits. Error-pair success requires accepting the original and rejecting the edit; valid-pair success requires accepting both.

		Qwen3-8B		Qwen3-14B	
Method	Input	Erroneous	Valid	Erroneous	Valid
CF-QK	Full solution	24/24	16/16	24/24	16/16
	Reasoning only	24/24	16/16	24/24	16/16
	Answer only	0/24	0/16	0/24	4/16
Verdict logits	Full solution	24/24	16/16	24/24	16/16
	Reasoning only	24/24	16/16	24/24	16/16

Results. On both models, CF-QK correctly separates all 24 erroneous pairs and accepts both solutions in all 16 valid pairs, with either the full solution or reasoning alone. Direct verdict logits attain the same results in these two conditions. With only the answer, CF-QK separates none of the erroneous pairs, as expected for identical paired inputs; it accepts both members of zero valid pairs on 8B and four on 14B. The contrast across representations shows that CF-QK responds to the supplied reasoning in this controlled test, even when the final answer remains correct.

C.3 ACTIVATION PATCHING

We patch Qwen3-8B activations on 80 counterfactual pairs, 40 from each dataset, using the counterfactuals from Appendix C.1 and both verdict-label mappings. Queries are captured at the colon in `Verdict:`; keys are captured at the terminal tokens of the two semantic verdict options. We measure the semantic next-token verdict-logit margin $m = z_{\text{correct}} - z_{\text{error}}$, averaged across mappings.

We fix 11 SA-selected heads and compare with 11 random heads from the final half of the 36-layer model, with at most one random head per layer. For K and Q+K controls, every query head uses its corresponding grouped-query-attention key head. We patch Q alone, K alone, or both, in each direction. Writing F for the intact run and C for the corrupted run, define

$$E = \tfrac{1}{2}\left[(m_{C\leftarrow F} - m_C) + (m_F - m_{F\leftarrow C})\right]. \tag{16}$$

Positive values indicate that intact activations increase the correct-verdict margin relative to corrupted activations, averaged over both intervention directions. Table 13 reports raw logit-margin changes, with 95% intervals from 10,000 bootstrap resamples.

Table 13: Bidirectional activation-patching effects on the semantic verdict-logit margin. "Difference" is selected minus random; intervals use 10,000 bootstrap resamples. Differences are computed before rounding.

Pair group	Patch	Selected	Random	Difference	95% CI
All 80	Q	0.0469	0.0039	0.0430	[0.0211, 0.0664]
All 80	K	0.2547	0.0359	0.2188	[0.1688, 0.2711]
All 80	Q+K	0.3352	0.0367	0.2984	[0.2426, 0.3590]
HLE, 40	Q	0.1164	0.0078	0.1086	[0.0750, 0.1414]
HLE, 40	K	0.2422	0.0281	0.2141	[0.1703, 0.2570]
HLE, 40	Q+K	0.4125	0.0344	0.3781	[0.3031, 0.4562]
MATH, 40	Q	-0.0227	0.0000	-0.0227	[-0.0328, -0.0125]
MATH, 40	K	0.2672	0.0437	0.2234	[0.1328, 0.3196]
MATH, 40	Q+K	0.2578	0.0391	0.2188	[0.1406, 0.3016]

K and joint Q+K patches have larger positive effects in the selected set, including on MATH where the final answer is unchanged; Q-only MATH patching has a negative effect. The random control uses the same number of heads, with at most one per layer; other activation statistics are not matched. An alternative head-selection/control protocol yields a negative overall Q-only selected-minus-control effect (-0.1557), further indicating dependence on the selected components and control set. These interventions show that the selected heads affect verdict logits in the tested examples.

C.4 HEAD SIGNALS AND COMPLEMENTARY ERRORS

Figure 3 examines Qwen3-8B readouts fitted with 50 source labels. Heads differ in both ranking performance and their errors on individual examples. The five-head SA ensemble attains 66.89% MATH BAcc, compared with 83.65% for the L2 readout. The error patterns motivate comparing aggregation rules. These fitted readouts are separate from the 11-head ensemble used for activation patching in Appendix C.3.

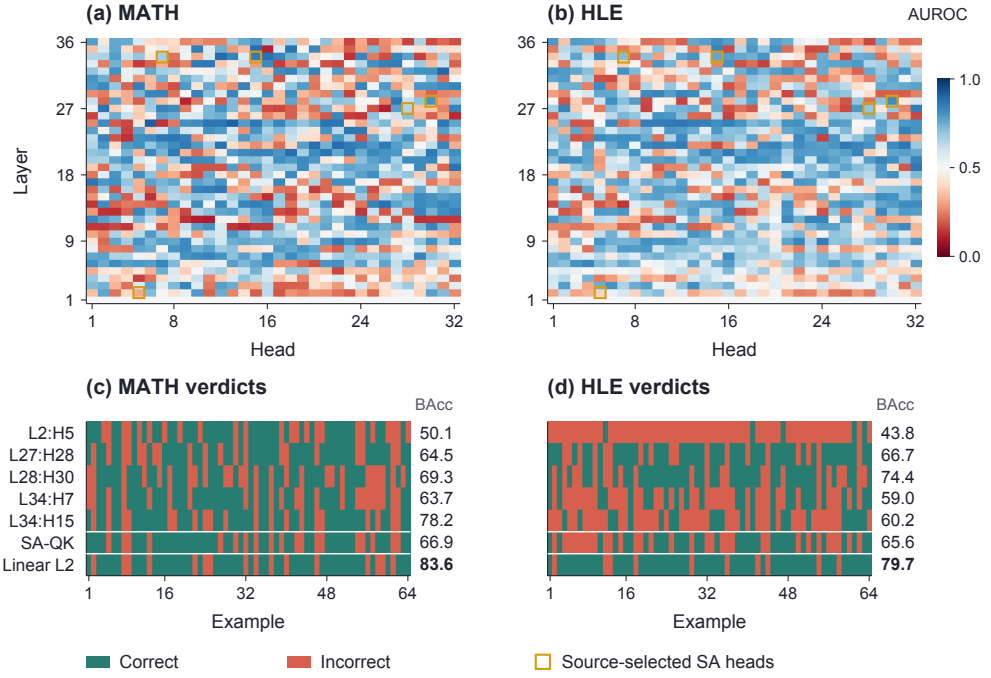


Figure 3: Head signals and verification errors in Qwen3-8B under the 50-label protocol. (a,b) Per-head AUROC on MATH (450 examples) and HLE (2,156), with score orientation fixed. Squares mark the five source-selected SA-QK heads. (c,d) Correct and incorrect verdicts on a fixed random sample of 64 examples per dataset. The final rows show SA-QK and the L2 readout over 128 heads. BAcc values refer to the complete evaluation sets. Layer and head indices are one-based.

C.5 REPEATED 50-LABEL CALIBRATION

Table 2 in the main text reports repeated 50-label calibration for four models using ten stratified splits and three search seeds per split, giving 120 model–split–seed configurations. We first average the three seeds within each split, then report the mean and SD over the ten split averages. The 450 MATH evaluation examples change between splits, but are shared by the compared methods within a split. The HLE evaluation set remains fixed. Thus MATH variation includes changes in both calibration and evaluation composition; HLE measures transfer variation induced by source calibration. These SDs are distinct from the conditional evaluation-bootstrap SDs in Table 1.

L2-QK has higher mean BAcc than calibrated verdict logits in all eight model–dataset comparisons, with gains of 4.06–16.05 percentage points. It also has the highest mean among the three methods in every comparison. Top-K averaging exceeds calibrated logits in seven of eight mean comparisons. On Qwen3-14B HLE, its higher mean corresponds to a win in three of ten splits; on Qwen3-32B HLE its mean is below the calibrated-logit baseline.

Paired improvements over calibrated logits. For each split, we subtract the calibrated verdict-logit BAcc from the QK BAcc and average the three search seeds. Table 14 reports the mean and SD of these paired differences over ten splits. A win is a split with a strictly positive seed-averaged difference. Paired SDs are computed from the differences, rather than from the marginal SDs in Table 2; means can differ by 0.01 points from subtraction of the rounded entries.

L2 retains a positive mean improvement on both MATH and HLE for every model, winning on eight to ten splits in each comparison. It improves on all ten splits for Qwen3-32B on both datasets and for DeepSeek-Llama-8B on HLE. Top-K wins on six to nine MATH splits and three to eight HLE splits. These comparisons use no HLE labels for calibration. Fitted thresholds also vary: the verdict-logit threshold is 5.900 ± 2.078 for Qwen3-8B and 6.691 ± 2.748 for Qwen3-14B, with mean and SD computed over all 30 fits per model.

Table 14: Paired BAcc improvements over calibrated verdict logits, in percentage points. Entries are mean $\pm$ SD over ten splits after averaging three search seeds within each split; wins count positive split-level differences. The models and evaluation sets match Table 2.

		Top-K QK		L2-QK	
Model	Dataset	ΔBAcc	Wins	ΔBAcc	Wins
Qwen3-8B	MATH	$+4.98 \pm 4.76$	9/10	$+6.40 \pm 3.67$	9/10
	HLE	$+14.64 \pm 11.83$	8/10	$+16.05 \pm 10.93$	8/10
Qwen3-14B	MATH	$+2.08 \pm 5.25$	7/10	$+4.67 \pm 4.26$	8/10
	HLE	$+1.55 \pm 10.96$	3/10	$+5.40 \pm 10.95$	8/10
Qwen3-32B	MATH	$+2.45 \pm 8.66$	6/10	$+9.06 \pm 2.36$	10/10
	HLE	-2.37 ± 13.05	5/10	$+5.00 \pm 6.61$	10/10
DeepSeek-Llama-8B	MATH	$+1.63 \pm 6.00$	8/10	$+4.06 \pm 4.84$	8/10
	HLE	$+7.98 \pm 13.55$	6/10	$+15.75 \pm 6.22$	10/10

Qwen3-32B threshold calibration. We reproduce the original Qwen3-32B calibration split through the repeated-calibration implementation and recover exactly 73.43% MATH and 68.19% HLE BAcc, as reported in Table 5. The feature-cache hashes, 128-head pool, normalization, regularization choice, fitted coefficients, decision threshold, and every evaluation score match the original fit. Table 2 averages fits on different calibration splits.

The original split selects a threshold of 0.3187 from ten threshold-calibration examples, producing many false-positive verdicts. Thresholds across the repeated fits have mean 0.4906 and SD 0.0939. To examine this effect while holding the fitted score fixed, we substitute 0.4906 for the original threshold; BAcc becomes 88.85% on MATH and 80.00% on HLE. This substitution is a post-hoc diagnostic and is not used in either result table. This diagnostic measures threshold sensitivity of the original classifier; the repeated-calibration results also reflect refitting on different source partitions.

C.6 SENSITIVITY TO THE CALIBRATION SPLIT

We vary the Qwen3-8B source calibration set while holding the same 100 MATH evaluation examples fixed. Figure 4 reports L2 performance across ten source splits at each label budget. These standard deviations measure variation in the fitted readout across splits, complementing the conditional evaluation uncertainty in Table 1. This study covers L2 for one model on a separate fixed evaluation set. Mean performance and candidate-pool overlap increase with the calibration budget, while single-head selection agreement remains low and is not monotonic. Candidate-pool overlap describes the heads admitted to the linear readout. The 45 pairwise comparisons share fitted models and are descriptive summaries, not independent observations.

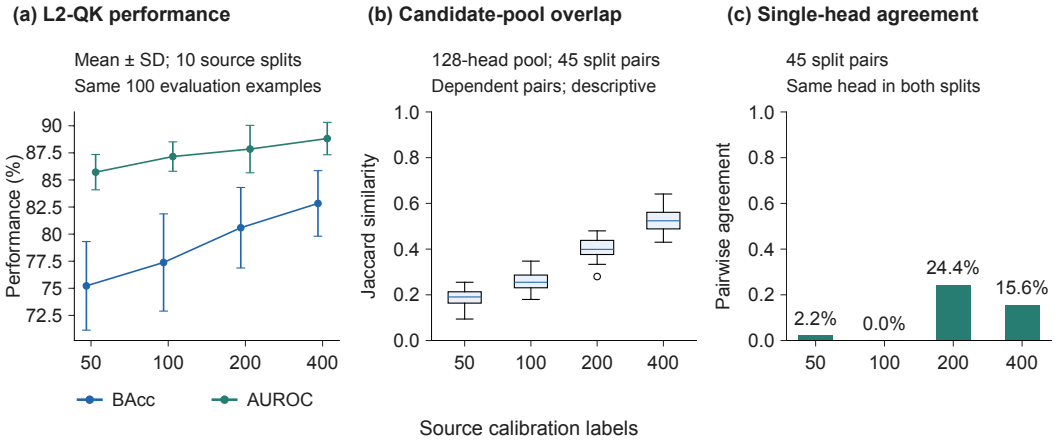


Figure 4: Calibration sensitivity for Qwen3-8B on a fixed 100-example MATH evaluation set. (a) L2-QK mean performance and SD across ten source splits. (b) Pairwise Jaccard similarity of the 128-head candidate pools used by L2. (c) Agreement of the selected single head across split pairs. Label budgets are 50, 100, 200, and 400; the 45 pairs in (b,c) are dependent.

C.7 Additive contributions to a linear verdict

The linear readout admits an exact decomposition of its decision score. For the fitted feature scale s_h, center μ_h, coefficient β_h, and intercept b, write

$$c_h(x) = \beta_h\big(m_h(x)/s_h - \mu_h\big), \qquad S(x) = b + \sum_h c_h(x), \tag{17}$$

where $m_h(x)$ is the semantic QK margin of head h. Figure 5 shows the first MATH evaluation example for the Qwen3-8B L2 readout fitted with 50 labels. It includes the eight largest absolute head contributions, the exact sum of the remaining 120, and the intercept. Their sum is 0.5638, above the fitted threshold 0.2639; the resulting correct verdict agrees with the answer label. This is an unmodified solution, separate from the counterfactual example in Figure 1. The decomposition gives the additive contributions to the fitted decision score.

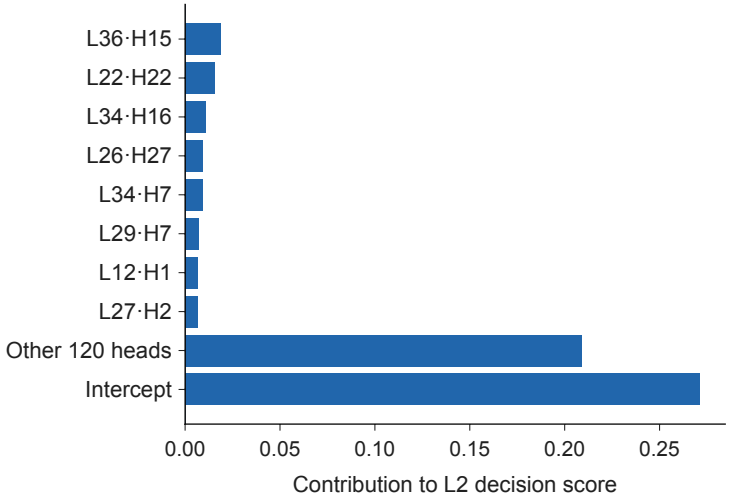


Figure 5: Exact additive decomposition of a Qwen3-8B L2 verification score on MATH. The 50-label fitted readout assigns a contribution to each of 128 heads. The displayed heads are ordered by absolute contribution; the remaining heads are summed without omission. The intercept is shown separately. Layer and head indices are one-based.

D Ensemble verification and scalar confidence baselines

D.1 Calibration, responsiveness, and option counterbalancing

Table 15 compares CF-QK, a stability-only alternative, SA-QK, and matched verdict-logit baselines across three models. SA-QK uses 400 source MATH labels; all other rows use zero correctness labels. The stability-only rule combines normalized Borda ranks using agreement across readout views, as specified in Appendix A. It uses all heads and computes its weights within each example. The responsive rule instead combines semantic sign votes from the final half of the layers and uses unlabeled batch statistics.

Stability across views can coexist with a nearly constant verdict. On Qwen3-14B/MATH, both stability-only variants classify every candidate as incorrect, yielding 37% accuracy and 50% BAcc. Responsive aggregation produces nonconstant predictions and reaches 81.02% BAcc. A similar result appears for DeepSeek-Llama-8B. The two-pass CF-QK variant improves DeepSeek-Llama-8B BAcc on both datasets, while its effect on Qwen depends on the model, dataset, and metric. The one-pass and two-pass rules are therefore reported separately throughout.

Table 15: Balanced accuracy of ensemble verification and aggregation variants (%). Qwen MATH evaluation uses 100 examples per model; DeepSeek-Llama-8B uses all 500. HLE uses 2,156 examples per model. Passes counts verifier forward passes per candidate; Labels counts source MATH labels. SA-QK transfers its selected heads, scales, and threshold unchanged to HLE. Responsive methods use all 500 unlabeled MATH examples or the full unlabeled HLE batch. The two-pass responsive method is the blend in Eq. (6); the stability-only two-pass rule is a separate aggregation procedure.

Method	Passes	Labels	MATH-500 BAcc	HLE-1/4 BAcc
Qwen3-8B				
Verdict logits	1	0	59.00	75.61
Verdict logits	2	0	61.32	79.87
Stability-only QK	1	0	70.81	77.46
Stability-only QK	2	0	50.00	49.88
CF-QK-R1P	1	0	74.05	80.84
CF-QK-RSym	2	0	67.65	78.21
SA-QK	1	400	80.72	79.43
Qwen3-14B				
Verdict logits	1	0	78.94	80.48
Verdict logits	2	0	78.94	79.76
Stability-only QK	1	0	50.00	51.72
Stability-only QK	2	0	50.00	50.75
CF-QK-R1P	1	0	81.02	74.79
CF-QK-RSym	2	0	81.02	77.20
SA-QK	1	400	89.27	65.44
DeepSeek-Llama-8B				
Verdict logits	1	0	62.53	68.26
Verdict logits	2	0	74.68	80.31
Stability-only QK	1	0	50.00	50.05
Stability-only QK	2	0	49.77	54.41
CF-QK-R1P	1	0	74.67	74.98
CF-QK-RSym	2	0	84.08	84.41

D.2 Scalar confidence and source-fitted thresholds

Table 16 compares indirect True/False logits and mean solution log probability for both Qwen models. The indirect score has a natural zero-threshold decision; raw log probability is evaluated as a ranking score. Source-fitted variants choose a threshold using the 400 non-test MATH examples and retain it for HLE. Threshold fitting changes the binary decision but preserves the score ordering and hence AUROC. Its effect on BAcc varies across domains: for example, fitting the indirect score improves Qwen3-14B MATH BAcc from 73.53% to 76.73%, but its frozen threshold reduces HLE BAcc from 77.66% to 67.74%. A useful ranking score can therefore perform poorly at a threshold fitted on another domain.

Table 16: Balanced accuracy of scalar confidence baselines (%). Labels counts source MATH calibration examples; no HLE labels are used. MATH evaluation uses the same 100-example holdout per model as the QK comparison, and HLE uses 2,156 examples.

Method	Labels	MATH-500 BAcc	HLE-1/4 BAcc
Qwen3-8B			
Indirect logits, zero threshold	0	58.71	77.56
Indirect logits, fitted threshold	400	67.71	69.28
Log probability, fitted threshold	400	50.78	49.11
Qwen3-14B			
Indirect logits, zero threshold	0	73.53	77.66
Indirect logits, fitted threshold	400	76.73	67.74
Log probability, fitted threshold	400	66.47	48.56

E MULTIPLE-CHOICE SELECTION ACROSS MODEL FAMILIES

E.1 CALIBRATION AND READOUT PROTOCOL

We evaluate calibrated single-head selection, following the option readout of Tulchinskii et al. (2024), across eight models, using both direct multiple-choice prompting and prompting with chain-of-thought (CoT) reasoning. The model receives all answer options in one prompt. For option j, the key is read at the end-of-line token following that option. The query is read at the colon in `ANSWER:` for direct prompting, or immediately before the final answer option after CoT generation. A head predicts the option with the largest query–key inner product. We evaluate every attention head on the calibration set, select the head with the highest accuracy, and apply it to disjoint evaluation questions. The output baseline uses the model's answer-option prediction. CoT generation is limited to 1,200 new tokens. Figure 7(b) illustrates the readout, and Figure 6 gives the prompt structure.

THINK LOGICALLY AND SELECT THE CORRECT ANSWER TO THE FOLLOWING QUESTION.
QUESTION: Colors in a soap bubble result from light
OPTIONS: A. dispersion. B. deflection. ...
With reasoning: Now think step by step and then give the final answer.
Direct prompting: `Answer:`

Figure 6: MCQA prompting scaffold. The reasoning condition appends the step-by-step instruction; direct prompting instead appends the answer prefix. The example abbreviates the option list for display. Both conditions include the complete question and choices.

For Qwen3, we distinguish the *projection readout*, which extracts Q/K before native per-head normalization, from the *native-normalized readout*. The Qwen3-8B MMLU-Pro results use the native-normalized readout; the Qwen3-14B and Qwen3-32B MMLU-Pro results and all Qwen3 HLE-MC results use the projection readout. Normalization changes relative QK scores and can affect which head is selected.

For each of five random splits, we sample 500 calibration questions with balanced correct-answer positions. Evaluation uses 2,000 MMLU-Pro questions (Wang et al., 2024) or 1,600 questions from a four-choice adaptation of Humanity's Last Exam (Phan et al., 2025). The latter MCQA collection contains 2,100 text-only questions: three incorrect choices are sampled from the original options for multiple-choice items, while distractors for open-ended items are generated by language models. Correct-answer positions are balanced. We denote this adaptation by HLE-MC; its chance accuracy is 25%. It is separate from the 2,156-example HLE-1/4 verification set (Appendix A).

We report answer accuracy (Acc.) and permutation accuracy (PA), following the joint-correctness criterion of Gupta et al. (2024). If I_i and I_i^π indicate a correct prediction on question i before and after randomly permuting its options, respectively, then

$$\mathrm{PA} = \frac{1}{N}\sum_{i=1}^{N} I_i I_i^\pi. \tag{18}$$

We retain the same calibration/evaluation splits. PA therefore measures the correctness of the predictions after reshuffle, but not the stability of the head selection procedure. Tables 17 and 19 report mean $\pm$ standard deviation across splits, except for the explicitly marked point estimates.

E.2 MMLU-PRO

The benefit of a single-head readout varies across models (Table 17). With direct prompting, Llama-3.1-8B improves from 16.4% to 32.9% accuracy and from 5.4% to 19.5% PA. The smaller DeepSeek-Qwen distillations also benefit, whereas the output baseline is stronger for several larger models. After CoT, QK accuracy is close to the output baseline for all eight models, with absolute differences of 0.4–0.9 percentage points. The reported standard deviations describe variation across splits. CoT improves many of the readouts, although the QK accuracy of Llama-3.1-8B is slightly lower with CoT than with direct prompting.

Table 17: Multiple-choice selection on MMLU-Pro (%). Entries are mean ± SD over five splits, each using 500 calibration and 2,000 evaluation questions. DeepSeek model names abbreviate `DeepSeek-R1-Distill-Qwen`. The Qwen3-8B rows use native Q/K normalization; Qwen3-14B and Qwen3-32B use the projection readout. PA repeats calibration after option permutation.

Model	Output Acc.	Output PA	QK Acc.	QK PA
Direct multiple-choice prompting				
Llama-3.1-8B	16.4 ± 0.5	5.4 ± 0.9	32.9 ± 0.7	19.5 ± 1.8
DeepSeek-Qwen-1.5B	13.2 ± 0.5	2.0 ± 0.3	18.1 ± 0.9	7.5 ± 1.7
DeepSeek-Qwen-7B	16.3 ± 0.4	4.6 ± 0.5	27.7 ± 0.5	13.2 ± 0.4
DeepSeek-Qwen-14B	40.9 ± 0.4	29.7 ± 0.3	44.0 ± 0.9	31.3 ± 1.1
DeepSeek-Qwen-32B	52.5 ± 0.6	41.1 ± 0.4	50.5 ± 1.5	38.2 ± 1.2
Qwen3-8B	45.3 ± 0.6	33.3 ± 0.6	45.5 ± 0.9	32.8 ± 0.7
Qwen3-14B	41.1 ± 0.3	28.3 ± 0.3	43.3 ± 1.2	30.8 ± 1.6
Qwen3-32B	47.9 ± 0.2	36.5 ± 0.2	42.6 ± 0.4	29.3 ± 2.7
Multiple-choice prompting with CoT				
Llama-3.1-8B	31.4 ± 0.4	20.3 ± 2.8	32.1 ± 1.2	21.0 ± 1.7
DeepSeek-Qwen-1.5B	26.9 ± 0.5	12.1 ± 0.4	27.3 ± 2.2	12.7 ± 1.2
DeepSeek-Qwen-7B	33.6 ± 0.4	20.1 ± 0.4	34.3 ± 0.3	20.8 ± 0.6
DeepSeek-Qwen-14B	51.4 ± 0.3	40.5 ± 0.4	52.2 ± 0.5	40.6 ± 0.9
DeepSeek-Qwen-32B	56.3 ± 0.6	44.3 ± 0.5	57.0 ± 0.7	45.0 ± 0.8
Qwen3-8B	53.9 ± 0.5	45.0 ± 0.4	54.7 ± 0.3	44.7 ± 0.7
Qwen3-14B	57.0 ± 0.5	46.6 ± 0.7	57.9 ± 0.5	45.8 ± 0.5
Qwen3-32B	60.4 ± 0.2	49.6 ± 0.2	59.9 ± 0.5	48.0 ± 0.5

Extraction-point comparison. Table 18 reports both Qwen3-8B extraction conventions. The native-normalized configuration gives higher accuracy and PA with direct prompting. Under CoT, accuracy is similar and PA is lower. The extraction point therefore affects the measured performance.

Table 18: Qwen3-8B MMLU-Pro QK readouts at two extraction points (%, mean ± SD across five splits). Both use 500 calibration and 2,000 evaluation questions. Projection denotes the readout before native Q/K normalization.

	Projection		Native-normalized	
Prompting	Acc.	PA	Acc.	PA
Direct	43.2 ± 0.6	26.0 ± 0.4	45.5 ± 0.9	32.8 ± 0.7
CoT	54.6 ± 0.6	45.8 ± 0.6	54.7 ± 0.3	44.7 ± 0.7

E.3 FOUR-CHOICE HLE (HLE-MC)

Table 19 extends the comparison to HLE-MC. Single-head QK selection improves direct-prompting accuracy for Llama-3.1-8B and the 1.5B and 7B DeepSeek-Qwen distillations. For the larger distillations and Qwen3 models, the output baseline is generally stronger. CoT does not yield a uniform increase in QK accuracy or PA. After CoT, Qwen3-14B and Qwen3-32B QK accuracy is within 0.45 percentage points of the output baseline, but PA is lower by 1.7 and 1.5 points, respectively.

Table 19: Multiple-choice selection on the 2,100-question HLE-MC adaptation (%). Each split uses 500 calibration and 1,600 evaluation questions. Entries are mean ± SD across five splits where available; daggered entries are reported point estimates without an SD. DeepSeek model names abbreviate `DeepSeek-R1-Distill-Qwen`. All Qwen3 rows use the projection readout before native Q/K normalization.

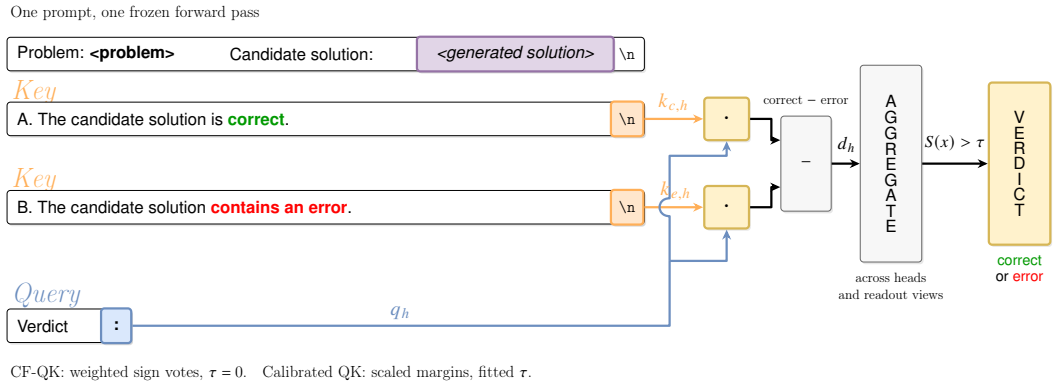
Model	Output Acc.	Output PA	QK Acc.	QK PA
Direct multiple-choice prompting				
Llama-3.1-8B	28.2 ± 0.4	10.6 ± 0.3	32.8 ± 1.2	13.4 ± 1.0
DeepSeek-Qwen-1.5B	27.7 ± 0.4	10.8 ± 0.3	32.4 ± 1.5	15.9 ± 0.3
DeepSeek-Qwen-7B	27.2 ± 0.6	12.0 ± 0.3	33.5 ± 2.0	14.9 ± 2.7
DeepSeek-Qwen-14B	36.1 ± 0.7	22.5 ± 0.4	34.5 ± 1.1	16.7 ± 1.7
DeepSeek-Qwen-32B	37.2 ± 0.5	25.6 ± 0.4	35.6 ± 1.0	24.8 ± 2.1
Qwen3-8B	35.5 ± 0.6	21.2 ± 0.5	35.8 ± 0.9	21.6 ± 1.7
Qwen3-14B	34.1 ± 0.5	20.6 ± 0.2	33.0 ± 0.6	19.0 ± 1.9
Qwen3-32B	36.8 ± 0.2	23.9 ± 0.3	32.4 ± 2.3	19.9 ± 1.5
Multiple-choice prompting with CoT				
Llama-3.1-8B	32.1 ± 1.0	14.2 ± 0.5	33.5 ± 0.4	15.2 ± 0.6
DeepSeek-Qwen-1.5B	29.6 ± 0.6	12.6 ± 0.4	33.1 ± 1.7	13.2 ± 1.3
DeepSeek-Qwen-7B	30.8 ± 0.2	15.2 ± 0.2	30.7 ± 1.1	13.7 ± 1.7
DeepSeek-Qwen-14B	34.6 ± 0.5	23.0 ± 0.4	32.0 ± 1.3	14.6 ± 1.3
DeepSeek-Qwen-32B	33.6 ± 0.9	16.2 ± 0.7	34.3 ± 1.8	20.3 ± 2.0
Qwen3-8B	33.4 ± 0.3	18.5 ± 0.4	35.3 ± 0.8	19.5 ± 2.0
Qwen3-14B	35.2 ± 0.2	20.9 ± 0.4	34.8 ± 2.1	19.2 ± 1.8
Qwen3-32B	35.3 ± 0.7	21.3 ± 0.3	34.8 ± 1.3	19.8 ± 1.4

E.4 Layerwise option selection

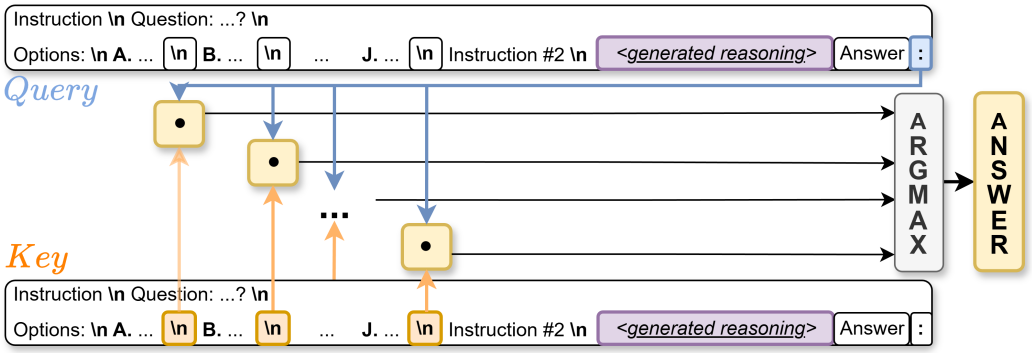
To locate useful option information without labeled head calibration, we choose, for each question and each layer, the head–option pair having the largest QK score. For Qwen3-8B on MMLU-Pro with CoT, layers 24, 25, and 28 achieve 52.5%, 54.9%, and 52.9% accuracy, compared with approximately 53.9% for the model. First-half layers are close to chance. Several later layers therefore approach the model's output accuracy. These results are descriptive: choosing the best layer from its evaluation accuracy would use labels for selection.

F QUERY–KEY READ POSITIONS

Figure 7 shows query and key positions for verification and multiple-choice selection. Panel (a) shows the one-pass verifier in Section 4: both verdict options precede the final query, and their head-wise scores are combined into a verification decision. Panel (b) shows the multiple-choice readout evaluated in Appendix E.



(a) One-pass verdict-option verification.



(b) Multiple-choice selection after reasoning.

Figure 7: Query–key positions for verification and multiple-choice selection. Blue marks queries and orange keys; dots denote inner products. (a) The canonical view reads the query at the colon in `Verdict:` and keys at the ends of the two options, using raw projections before Q/K normalization and RoPE. CF-QK also combines nearby read anchors. Correct-minus-error margins feed the method-specific aggregation rule, with a zero threshold for CF-QK and a source-fitted threshold for calibrated readouts. The prompt is abbreviated. (b) The answer-boundary query is compared with keys following the options; direct prompting omits the second instruction and generated reasoning.

G DIFFERENCES FROM PRIOR QUERY–KEY READOUTS

Attention products of a frozen model have been read out for multiple-choice selection and logical-consistency classification (Tulchinskii et al., 2024; 2025), document re-ranking (Chen et al., 2025; Zhang et al., 2025b), and list filtering (Sen Sharma et al., 2026). The verifier in this paper differs from these readouts in five respects.

- **Object.** The judged input is a complete reasoning trace that the model generated itself for an open-ended problem, scored without a reference answer, rather than options or statements supplied in the prompt (Section 4).
- **Readout.** Each head scores a signed contrast between two explicit verdict keys, $q^\top(k_c - k_e)$, over nine anchor views (Eq. 2). Reading the product before rotary encoding keeps a positional rotation, which would differ between the two option keys, out of the contrast, and the signed votes of different heads are commensurable, which is what makes aggregation possible.
- **Decision without labels.** CF-QK weights all final-half heads by agreement across views, coverage, and responsiveness computed on an unlabeled batch, and decides at a fixed zero threshold. No correctness label enters head selection, weighting, or thresholding, and responsiveness acts as a weight over all heads rather than as a criterion for choosing one (Section 4.2; Tables 1, 6, and 8).
- **Fifty labels.** The calibrated readouts fit head selection, a linear classifier, or subset search on 30/10/10 examples, repeated over ten splits, and transfer with frozen parameters (Tables 2 and 14).
- **Evidence about the signal.** Fixed-answer edits separate reasoning from answer content, matched counterfactuals change one step, and query/key patching in the selected heads moves the model's verdict margin relative to random heads (Sections 7.3 and 7.4).

Comparisons with verdict logits, indirect confidence, solution likelihood, an open-generation judge, and a process reward model score the same candidates and, where applicable, use the same labels and verifier passes (Tables 5, 15, and 10).