FROM ROUTING DIFFERENCES TO CANDIDATE SELECTION IN MOE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixture-of-Experts (MoE) routing carries meaningful structure and is increasingly treated as an actionable interface. Targeted expert activation or deactivation can steer model behavior, motivating us to ask whether naturally occurring routing variation also reflects answer correctness. Under stochastic sampling, the same question can yield both correct and incorrect solutions, creating a practical need to identify which candidates are worth selecting or retaining. Using a paired-trajectory design, we compare correct and incorrect candidates for the same question across five MoE models and five reasoning datasets, connecting routing differences, cross-question readout, and candidate selection. Routing differences exceed a matched random-label baseline in 21 of 24 settings after Holm correction. Supervised probes recover correctness-related information on unseen questions, with readout strength varying across models and tasks. Adding routing to candidate scoring improves both Top-1 accuracy and Coverage@2 for four of five DAPO-Math models, including a 7.3 percentage-point Coverage@2 gain for GLM. Retention gains broadly follow relative probe advantages across models and tasks. Mean routing distance offers little additional benefit over text diversity or existing scores and lowers rare-correct retention in all nine evaluated settings. These results establish routing as a source of learnable candidate quality information whose selection value depends on both signal strength and how it is used.

1 INTRODUCTION

Mixture-of-Experts (MoE) has become a widely used architecture for large reasoning models (Cai et al., 2025), including the DeepSeek (DeepSeek-AI, 2025), Qwen (Yang et al., 2025), and GLM (Team, 2025) families. At each MoE layer, the router selects the experts that process each token, providing an observable view of expert allocation. Recent work has identified stable cross-layer routing patterns (Li et al., 2026) and language-dependent expert allocation (Bandarkar et al., 2026). Safety studies further identify experts with concentrated roles in safety behavior (Lai et al., 2025) and show the value of preserving safety-aligned routing during fine-tuning (Kim et al., 2026).

This structure also proves actionable. Targeted expert activation or deactivation can steer safety properties (Fayyaz et al., 2026), while amplifying a small set of cognitive experts at inference time can steer reasoning behavior (Wang et al., 2025). These successes motivate examining whether naturally occurring routing variation also carries information about the correctness of generated answers.

Under stochastic sampling, the same question routinely produces both correct and incorrect solutions. When several candidates are available, we may select one answer or retain a few for further verification (Snell et al., 2025). **Can routing help distinguish and select correct answers to the same question?**

We use a paired-trajectory design to compare correct and incorrect candidates while holding the input fixed. We first measure routing differences against a matched random-label baseline, then train supervised probes to distinguish correct from incorrect candidates on held-out questions. We next evaluate two uses of routing for selection, learning candidate scores and selecting diverse candidate sets from routing distances (Figure 1). We study five MoE models from the Qwen (Yang et al., 2025), Gemma (Farabet & Lacombe, 2026), and GLM (Team, 2025) families. The five reasoning datasets span mathematics (Cobbe et al., 2021; Yu et al., 2025; Lightman et al., 2024), question answering (Yang et al., 2018), and code (Chen et al., 2021).

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

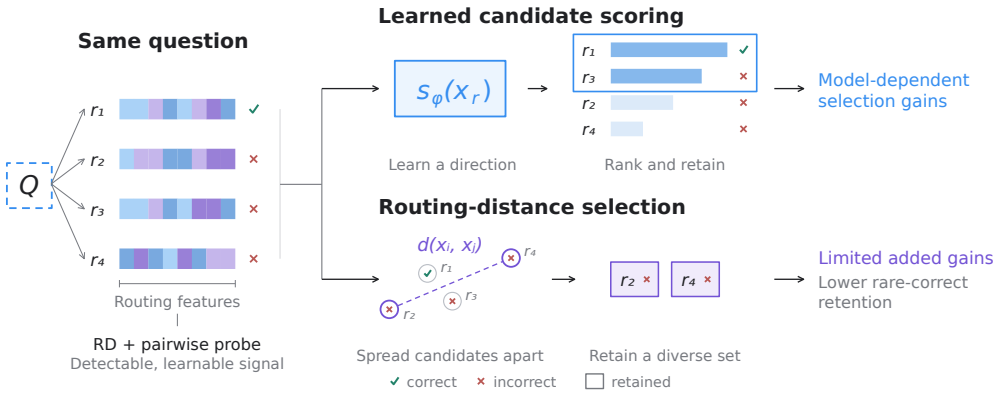


Figure 1: **From routing differences to candidate selection.** Same-question candidates connect outcome-associated differences, cross-question readout, and selection through learned scores or routing distance.

Correct and incorrect candidates exhibit routing differences beyond the matched null in 21 of 24 settings after correction. Linear routing probes achieve higher AUC than length-only probes on all five models in the common DAPO-Math comparison. These results show that routing contains quality signals that can be learned across questions, with their strength varying across models and tasks.

Candidate scoring turns these signals into practical selection gains. Adding routing features improves both Top-1 accuracy and Coverage@2 for four of the five DAPO-Math models, including a 7.3 percentage-point Coverage@2 gain for GLM. Larger retention gains broadly accompany stronger probe advantages, with corresponding changes for GLM and Qwen3-30B across DAPO-Math and HotpotQA. Mean routing distance offers little additional benefit over text diversity or existing scores and lowers rare-correct retention in all nine evaluated settings. Together, these findings show that routing's selection value depends on both signal strength and how the signal is used.

Contributions.

- An **outcome-focused routing evaluation** that compares correct and incorrect candidates for the same question and reveals routing differences across models and tasks through matched-null calibration.
- An **experimental connection between routing readout and selection gains**, showing how learned scoring uses routing information and how its gains broadly follow signal strength across models and tasks.
- An **empirical comparison of candidate scoring and routing diversity** that identifies their different contributions to selection and provides guidance for using routing information.

2 Comparing Candidates for the Same Question

2.1 Same question, different outcomes

For each question q, let $\mathcal{R}_q$ denote the sampled candidate rollouts with recorded routing and correctness labels $c_r \in \{0, 1\}$, where $c_r = 1$ denotes a correct answer. A question is *mixed* when $\mathcal{R}_q$ contains both correct and incorrect candidates, allowing outcome comparisons while holding the question fixed. Probes and learned scorers use correctness labels from training questions and are evaluated on disjoint test questions, with all candidates from a question kept together. A separate controlled experiment constructs noisy synthetic scores from correctness labels to compare selection rules under a fixed score provider (Appendix D.7).

2.2 Routing features across generation progress

For token t of rollout r, let $S_l^{(r,t)}$ contain the K selected experts at MoE layer $l \in \{1, \dots, L\}$, drawn from E routed experts. We exclude always-active shared experts. We group paragraph-based segments of the recorded thinking span into early, middle, and late bins by relative position, indexed by $\phi \in \Phi = \{1, 2, 3\}$. Let $I_\phi^{(r)}$ be the set of token positions in phase ϕ. The normalized expert-selection frequency is

$$p_{l,\phi,e}^{(r)} = \frac{1}{K|I_\phi^{(r)}|} \sum_{t \in I_\phi^{(r)}} \mathbf{1}\{e \in S_l^{(r,t)}\}, \qquad \sum_{e=1}^{E} p_{l,\phi,e}^{(r)} = 1. \tag{1}$$

We write $p_{l,\phi}^{(r)}$ for the corresponding expert distribution. Stacking these distributions gives a phase $\times$ layer $\times$ expert occupancy tensor, whose flattened vector is $x_r \in \mathbb{R}^{|\Phi|LE}$. This representation preserves phase, layer, and expert identity while averaging over token order within each phase. The same representation supports the outcome analyses and selection experiments. Segmentation and aggregation details appear in Appendix B.

2.3 Models, data, and evaluation

We study five MoE models from the Qwen (Yang et al., 2025), Gemma (Farabet & Lacombe, 2026), and GLM (Team, 2025) families. They are Qwen3-30B-A3B, Qwen3.5-35B-A3B, Qwen3-235B-A22B, Gemma-4-26B-A4B-it, and GLM-4.7-Flash, abbreviated in tables as Qwen3-30B, Qwen3.5, Qwen3-235B, Gemma, and GLM. The five datasets are DAPO-Math (Yu et al., 2025), GSM8K (Cobbe et al., 2021), MATH-500 (Lightman et al., 2024), HotpotQA (Yang et al., 2018), and HumanEval (Chen et al., 2021). DAPO-Math is the common task for the main cross-model comparison. Its subset was screened for mixed outcomes under Qwen3-30B-A3B before the subsequent evaluation, so the analysis is conditional on this construction. The outcome evaluation samples 16 rollouts per question at temperature 0.7 and captures expert assignments by replaying saved trajectories.

Routing-difference and probe analyses cover 24 available model–dataset settings. Learned selection covers all five models on DAPO-Math and HotpotQA, plus GLM on HumanEval. Each experiment reports its eligible question counts. Uncertainty is estimated by resampling questions, and method differences use paired comparisons on the same questions. Appendix A gives model specifications, dataset construction, and generation settings. We also provide interpretive observations in Appendices E and F.

3 Do Correct and Incorrect Answers Route Differently?

Candidates sharing a question allow us to compare expert allocation while holding the input fixed.

Measuring routing differences. For each mixed question q, let n_q^+ and n_q^- count the correct and incorrect candidates in $\mathcal{R}_q$. Balanced subsampling draws $m_q = \min(n_q^+, n_q^-)$ candidates from each outcome group. Let $\bar{p}_{q,l,\phi}^+$ and $\bar{p}_{q,l,\phi}^-$ denote their mean expert distributions, averaged over the balanced draws. We measure routing difference (RD) by total variation,

$$\mathrm{RD}_{q,l,\phi} = \frac{1}{2} \sum_{e=1}^{E} \left| \bar{p}_{q,l,\phi,e}^+ - \bar{p}_{q,l,\phi,e}^- \right|. \tag{2}$$

RD lies in $[0, 1]$, with larger values indicating greater separation between the group means. We average RD over layers and phases, and over questions with weights m_q. Algorithm 1 in Appendix B gives the complete estimation procedure.

Matched permutation baseline. Within-question label permutations preserve outcome counts and candidate pools, providing a matched baseline for finite-sample differences. We repeat the estimator for 1,000 permutations and apply Holm correction across 24 settings (Appendix B.2).

Table 1: **Outcome-related routing differences.** Excess is observed minus matched-null RD, using the same estimator; $*$ marks Holm-adjusted $p < 0.05$. Rows cover all DAPO models and representative cross-task settings. Appendix B.2 reports all 24 settings and intervals.

Model / task	Q	Observed RD	Null RD	Excess	
DAPO-Math, all models					
Qwen3-30B	247	0.0657	0.0496	+0.0161	$*$
Qwen3.5	32	0.1062	0.0858	+0.0204	$*$
Gemma	75	0.0466	0.0405	+0.0061	$*$
GLM	150	0.0793	0.0681	+0.0112	$*$
Qwen3-235B	158	0.0657	0.0519	+0.0138	$*$
Cross-task comparisons					
GLM × GSM8K	38	0.1256	0.0962	+0.0294	$*$
Qwen3-30B × HotpotQA	62	0.0833	0.0628	+0.0205	$*$
GLM × HotpotQA	110	0.0960	0.0807	+0.0153	$*$
GLM × HumanEval	46	0.1394	0.1219	+0.0174	$*$
GLM × MATH-500	31	0.1228	0.1037	+0.0191	$*$

Outcome-related routing differences. Correct and incorrect candidates exhibit routing differences across models and tasks (Table 1). Observed RD exceeds the matched null in 23 of 24 settings, with significant differences in 21 after Holm correction, spanning mathematics, question answering, and code. These results show that expert allocation carries information about answer correctness even when the input question is held fixed.

Length control. Routing differences remain at a similar scale after matching correct and incorrect candidates by rollout length across all five DAPO models (Figure 2). For Qwen3-30B, RD changes from 0.0657 to 0.0703, with 245 of 247 questions supporting the matched comparison. The persistence of these differences among candidates of comparable length supports treating expert allocation as an outcome-related feature. Appendix B.3 reports the matching support and same-label references.

We next ask whether this information can be learned to distinguish correct and incorrect candidates on unseen questions.

4 Can Routing Distinguish Correct and Incorrect Candidates?

We test whether a probe trained on one set of questions can distinguish correct and incorrect candidates on held-out questions from the same model–dataset setting.

Learning across questions. Each correct–incorrect pair (r_a, r_b) from $\mathcal{R}_q$ contributes the feature difference $x_{r_a} - x_{r_b}$, with both pair orientations included. An L_2-regularized logistic probe assigns the pair score

$$f(r_a, r_b) = w^\top (x_{r_a} - x_{r_b}) + b. \tag{3}$$

Feature scaling and hyperparameter selection use training folds only, and all candidates of a question remain in one fold. The length baseline uses the difference in log-transformed rollout lengths under the same splits. Pair-order AUC is computed within each test question and averaged equally across questions. For a linear probe, w defines a scoring direction $s_{\text{probe}}(r) = w^\top x_r$, connecting the readout to candidate judgment.

Readout across models. Routing probes have higher AUC than the length baseline on all five DAPO models, with a median advantage of 9.1 percentage points (Table 2). Across the full matrix, 17 of 24 differences are positive. GLM and Qwen3.5 show clear relative advantages, while the DAPO increments for Qwen3-30B and Qwen3-235B are smaller. These results show that expert allocation contains quality information that can be learned across questions, with differing discriminative value across models.

Table 2: **Readout varies across models and tasks.** Pair-order AUC (%) has chance level 50. Differences are routing minus length (pp), with paired 95% question-bootstrap intervals. Rows match Table 1; full results appear in Appendix C.

Model / task	Q	Length	Routing	Difference	95% CI
DAPO-Math, all models					
Qwen3-30B	247	57.36	57.74	+0.38	[−5.64, +6.63]
Qwen3.5	32	37.91	57.03	+19.12	[+0.58, +39.21]
Gemma	75	40.92	50.00	+9.08	[−1.75, +19.08]
GLM	150	50.33	59.50	+9.17	[+1.33, +17.54]
Qwen3-235B	158	57.20	58.62	+1.42	[−5.38, +8.67]
Cross-task comparisons					
GLM × GSM8K	38	72.86	74.51	+1.64	[−9.21, +11.68]
Qwen3-30B × HotpotQA	62	43.85	57.26	+13.41	[−1.92, +29.44]
GLM × HotpotQA	110	50.23	51.14	+0.91	[−7.67, +9.72]
GLM × HumanEval	46	45.79	61.55	+15.76	[−0.89, +29.56]
GLM × MATH-500	31	53.43	72.58	+19.15	[+5.04, +34.27]

Task dependence. The readout also changes with the task. GLM has a larger advantage over length on DAPO-Math than on HotpotQA, while Qwen3-30B exhibits the reverse pattern. Their relative advantages on HotpotQA are 0.9 and 13.4 percentage points, respectively.

Sensitivity to progress bins. The readout pattern is broadly preserved when equal-token thirds replace paragraph-based progress bins (Figure 3). Keeping the trajectories, labels, and question folds fixed, routing AUC changes by at most 2.3 percentage points across the three available DAPO comparisons. This stability supports using the occupancy representation to summarize routing across generation progress without relying on one choice of boundaries. Appendix B.4 gives the paired estimates and intervals.

We next test whether routing improves candidate selection and how the gains relate to these differences in readout strength.

5 How Does Routing Help Candidate Selection?

We test whether routing helps select or retain correct answers under a fixed candidate budget. Top-1 selection returns one answer, while Coverage@2 measures whether two retained candidates include a correct answer for later checking. Rare-retain@2 asks whether the only correct answer in a candidate pool survives selection.

5.1 Learned Candidate Scoring

Scorers and evaluation. We compare a base scorer S_B with a routing-augmented scorer S_{BR} using the same training folds and test questions. Both are linear scorers trained with candidate-correctness supervision. S_B uses rollout length and four simple text statistics, comprising log character count, log token count, newline frequency, and digit frequency. S_{BR} adds the routing feature vector x_r defined in Section 2.2. The probe in Section 4 learns a paired judgment direction, while these scorers learn candidate-level correctness scores from the same representation. We rank candidates by their held-out scores and evaluate Top-1 and Coverage@2. The paired question-bootstrap intervals use 10,000 resamples and are exploratory, without multiplicity correction (Appendix D.1).

Gains from routing scores. Adding routing improves both Top-1 and Coverage@2 for four of the five DAPO models (Table 3). GLM's Coverage@2 rises from 76.7% to 84.0%, while Qwen3.5 gains 9.4 percentage points. For GLM, the same budget of two candidates retains a correct answer for roughly seven more questions per hundred. The readable quality signal therefore translates into a concrete selection benefit.

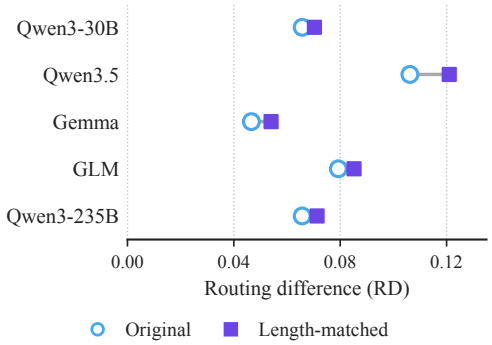


Figure 2: **RD persists after length matching.** Original and length-matched RD on five DAPO models. Matching support and estimates appear in Appendix B.3.

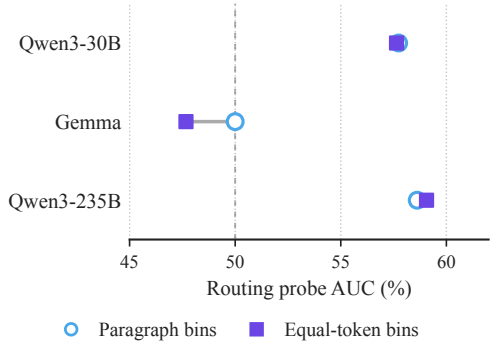


Figure 3: **Readout is stable across progress bins.** Paragraph versus equal-token bins in three DAPO settings, with fixed trajectories and folds. Dashed line marks chance. Intervals appear in Appendix B.4.

Table 3: **Routing scores improve retention in most settings.** Probe advantage is routing minus length pair-order AUC (pp). Arrows show $S_B \rightarrow S_{BR}$ (%); ΔCoverage@2 is in pp with paired 95% intervals. Green/red shading marks positive/negative estimates, with intensity indicating magnitude. All 11 settings are shown; the pooled row weights DAPO models equally.

Model / task	Q	Probe adv.	Top-1	Coverage@2	ΔCoverage@2 [95% CI]
Panel A. DAPO-Math					
Qwen3-30B	247	+0.38	68.02 → 71.66	82.59 → 85.02	+2.43 [−3.24, +8.10]
Qwen3.5	32	+19.12	46.88 → 59.38	75.00 → 84.38	+9.38 [−9.38, +28.12]
Gemma	75	+9.08	60.00 → 58.67	72.00 → 70.67	−1.33 [−13.33, +10.67]
GLM	150	+9.17	63.33 → 72.00	76.67 → 84.00	+7.33 [+0.67, +14.00]
Qwen3-235B	158	+1.42	66.46 → 67.72	80.38 → 81.01	+0.63 [−5.70, +6.96]
Pooled, equal weight	662	–	–	–	+3.69 [−1.43, +8.75]
Panel B. Cross-task comparisons					
Qwen3-30B × HotpotQA	62	+13.41	45.16 → 46.77	58.06 → 67.74	+9.68 [−3.23, +22.58]
GLM × HotpotQA	110	+0.91	55.45 → 64.55	68.18 → 70.00	+1.82 [−6.36, +10.00]
GLM × HumanEval	46	+15.76	86.96 → 84.78	100.00 → 95.65	−4.35 [−10.87, +0.00]
Qwen3-235B × HotpotQA	123	−1.70	58.54 → 56.10	70.73 → 72.36	+1.63 [−7.32, +10.57]
Gemma × HotpotQA	55	+3.58	61.82 → 47.27	69.09 → 63.64	−5.45 [−16.36, +5.45]
Qwen3.5 × HotpotQA	48	−4.17	50.00 → 50.00	77.08 → 68.75	−8.33 [−22.92, +6.25]

Signal strength and task dependence. The distribution of retention gains broadly follows the probe advantages (Figure 4). On DAPO, GLM and Qwen3.5 have both clear relative readout advantages and larger selection gains, while Qwen3-30B and Qwen3-235B gain less. On HotpotQA, Qwen3-30B has the larger probe advantage and gains 9.7 points in Coverage@2, compared with 1.8 points for GLM. Their relative positions change together across tasks, supporting the use of task-specific signal measurements to understand retention gains.

Across the five models, four show same-direction changes in probe advantage and Coverage@2 gain between DAPO-Math and HotpotQA. Qwen3.5 and Gemma also show lower values for both on HotpotQA. This broader pattern reinforces the task dependence of routing's selection value and motivates evaluating its quality signal in the model–task setting where it will be used.

Across all 11 settings, seven have positive Coverage@2 increments. GLM on HumanEval shows the reverse outcome, with base coverage already at 100% and routing-augmented coverage falling to 95.7%, alongside a decline in candidate-ranking AUC (Appendix D.2). Routing's practical value thus also depends on what it adds to the available scoring features.

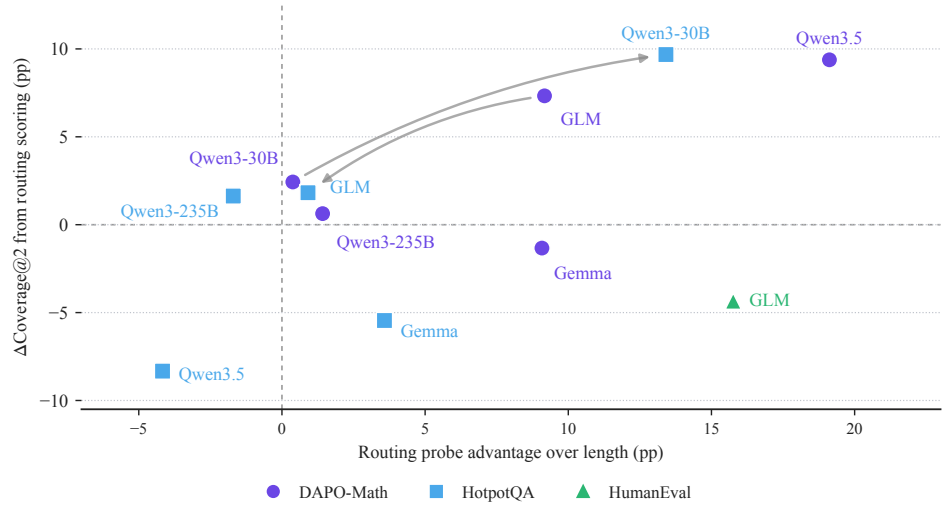


Figure 4: **Selection gains broadly follow readout strength.** Routing's AUC advantage over length versus its Coverage@2 gain (both in pp), across all 11 settings in Table 3. Arrows track Qwen3-30B and GLM from DAPO-Math to HotpotQA.

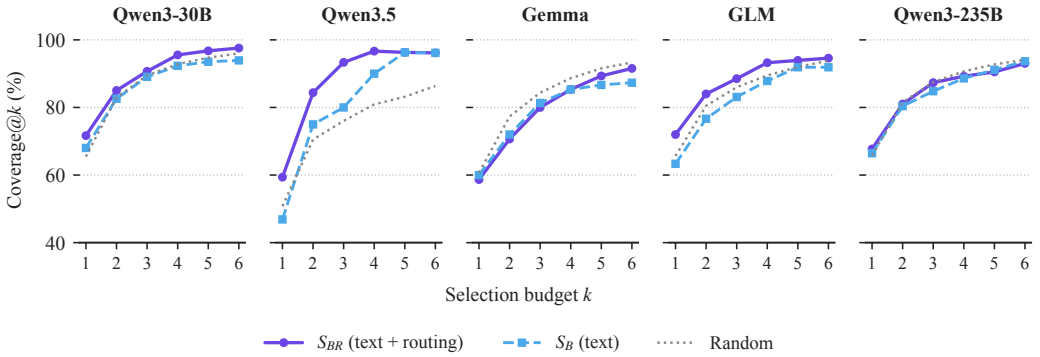


Figure 5: **Scoring gains across DAPO selection budgets.** GLM gains across budgets, while Qwen3.5 gains most at small budgets. Solid/dashed curves denote S_{BR}/S_B; dotted curves give random-selection expectations. Each budget k includes questions with at least k candidates.

Table 4: **Routing distance adds little beyond text and scores.** Coverage differences are in pp. Coverage@4 gives medians; $P_2 - P_1$ gives the equal-weight DAPO mean [95% interval]. Rare-retain requires ≥ 5 rare-correct questions per setting. Protocols and full results appear in Appendix D.

Comparison	Metric	Settings	Result
Routing-Diverse − Random	Coverage@4	24	+2.47
Routing-Diverse − Text-Diverse	Coverage@4	24	−0.32
Score+routing − Score-Only	Coverage@4	24	−0.22
$P_2 - P_1$	Coverage@2	5	−2.07 [−4.42, −0.13]
Score+routing − Score-Only	Rare-retain@2	9	Decreases in 9/9

Selection budgets. The budget sweep in Figure 5 shows that GLM's advantage persists across the displayed budgets. Qwen3.5 gains most when only a few candidates can be retained, and the two scorers converge as the budget grows. Learning a routing score can therefore help retain the correct answer earlier under a limited checking budget. Cross-task budget curves appear in Appendix D.5.

Robustness across models. The positive DAPO scoring trend persists under alternative model weights and after leaving out any one model. The average Coverage@2 gain is 3.7 points with equal model weights and 3.0 points with question weights. All five leave-one-model-out averages remain positive, showing that the aggregate improvement is shared across models (Appendix D.4).

5.2 Selecting with Routing Distance

We evaluate mean routing distance in both rule comparisons and learned pair models.

Diversity rules. Full-pool rules compare routing diversity with random selection and text diversity. Text distance uses token 3-gram Jaccard distance, while routing distance averages Jensen–Shannon distances across layers and phases. A controlled shortlist comparison uses synthetic scores formed by adding Gaussian noise to correctness labels. Given the same synthetic scores and top-six shortlist, Score-Only takes the top k, while Score+routing keeps the highest-scoring candidate and fills the remaining slots by routing diversity.

Pairwise selection objective. Retaining a correct answer in a pair is equivalent to avoiding joint failure. For a pair $A = \{r_a, r_b\}$ from the eligible pool $\mathcal{P}_q \subseteq \mathcal{R}_q$, let $\widehat{p}_{00}(A)$ estimate the probability that $c_{r_a} = c_{r_b} = 0$ from the pair's features. The pair selector maximizes predicted coverage by choosing

$$\widehat{A}_q = \arg \min_{\substack{A \subseteq \mathcal{P}_q \\ |A|=2}} \widehat{p}_{00}(A), \qquad \widehat{\text{Coverage}}(A) = 1 - \widehat{p}_{00}(A). \tag{4}$$

We learn joint failure directly from pair labels, allowing the selector to account for how two candidates fail together. P_0 uses base scores, lengths, and text distance, P_1 adds routing candidate scores, and P_2 adds mean routing distance. Thus $P_2 - P_1$ tests whether routing distance helps avoid joint failure beyond the available candidate quality scores and text features. Appendix D specifies the rules, pair models, and score provider.

Incremental value of routing distance. Mean routing distance does not show the gain pattern of learned scoring (Table 4). Routing diversity improves over random selection, but performs comparably to or slightly below text diversity and existing scores. The learned pair comparison also has a negative average distance increment on DAPO, reducing Coverage@2 by about 2.1 points under equal weighting. The increment remains negative with question weights and in every leave-one-model-out average (Appendix D.4). Increasing routing separation between candidates therefore offers little consistent additional selection value.

Rare-correct retention. The rare-correct comparison makes the selection cost more direct. Rare-retain@2 decreases in all nine settings, with the largest drop of 10.7 percentage points (Figure 6). The tested distance rule can therefore discard correct candidates that would otherwise be retained. How different two candidates are in routing does not directly determine which answers are worth keeping.

Takeaway. Learning candidate scores connects routing features to answer correctness, allowing readable quality information to guide selection. The gains broadly accompany stronger relative readouts, while mean routing distance adds little beyond scores and text information and reduces rare-correct retention in the tested settings. These comparisons favor learning how expert allocation relates to candidate quality and using that relation to retain correct answers under a fixed budget.

6 Related Work

Routing structure and expert specialization. Sparse MoE layers scale capacity by activating a subset of experts per token (Shazeer et al., 2017; Fedus et al., 2022). Analyses identify persistent cross-layer routing patterns (Li et al., 2026) and language-dependent allocation (Bandarkar et al., 2026). Training methods encourage router–expert alignment (Lv et al., 2026), expert specialization (Guo

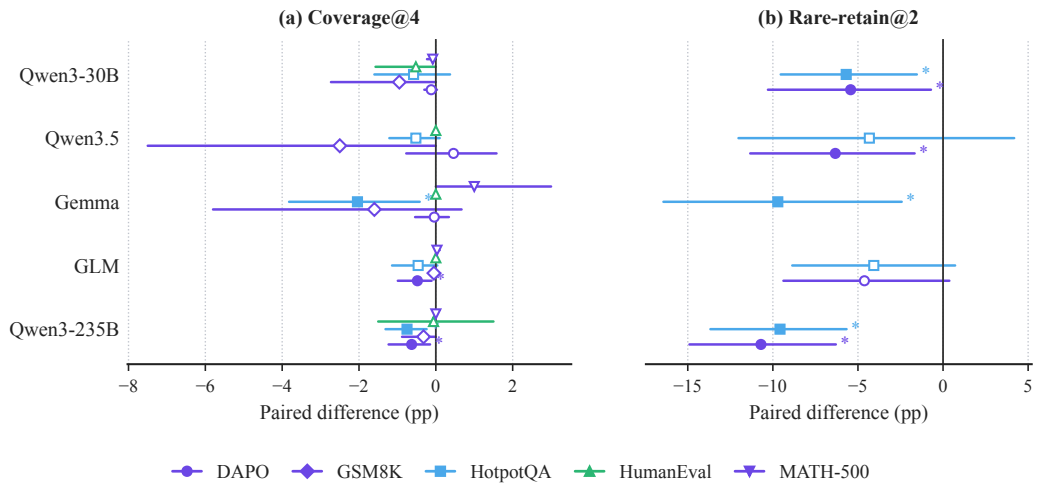


Figure 6: **Routing diversity adds little coverage and lowers rare-correct retention.** Synthetic-score comparisons show (a) Coverage@4, with score-only near ceiling, and (b) Rare-retain@2 for settings with at least five rare-correct questions. Bars show paired 95% question-bootstrap intervals; open markers include zero and asterisks exclude it.

et al., 2025), and denser router gradients (Panda et al., 2025). We examine whether its natural variation also distinguishes correct and incorrect candidates for the same question.

Routing intervention and reasoning. SteerMoE identifies behavior-associated experts and changes their activation to steer safety and faithfulness (Fayyaz et al., 2026). RICE reinforces experts associated with cognitive operations to improve reasoning (Wang et al., 2025). These interventions demonstrate that expert allocation can affect behavior. We read routing from naturally sampled trajectories and evaluate the information it provides for judging and selecting the resulting candidates.

Candidate scoring and diversity selection. Learned verifiers rank candidate solutions using outcome or process supervision (Cobbe et al., 2021; Lightman et al., 2024). Test-time scaling studies further examine how verification and search allocate inference computation (Snell et al., 2025). We study routing as an internal source of candidate quality information, evaluating both learned scores and distance-based set selection. Our study connects same-question routing differences, cross-question prediction, and candidate selection to examine how routing's value varies with the model, task, and way it is used.

7 Conclusion

Routing contains broadly observed information about answer correctness that can be read out on unseen questions. Learning candidate scores from this information improves selection in most evaluated settings, with retention gains broadly following the relative readout strength across models and tasks. Mean routing distance offers little consistent additional selection value and reduces retention when correct answers are rare. These findings support treating routing as a candidate quality feature, learning its relation to correctness, and evaluating its value through the intended selection objective.

Limitations. Our experiments focus on occupancy representations and linear scoring within mixed-question candidate pools. Future work can extend this framework to richer routing representations and broader candidate populations, and study how learned routing scores transfer across tasks. Combining routing scores with learned verifiers or reward models under a shared selection budget offers a further direction for using routing information in test-time inference.

AI USE STATEMENT

We used generative AI tools to assist with experimental code and manuscript editing. The authors reviewed the assisted work and take responsibility for the research design, analysis, and final content.

REFERENCES

Lucas Bandarkar, Chenyuan Yang, Mohsen Fayyaz, Junlin Hu, and Nanyun Peng. Multilingual routing in mixture-of-experts. In *The Fourteenth International Conference on Learning Representations*, 2026.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, 37 (7):3896–3915, 2025.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

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

DeepSeek-AI. DeepSeek-V3 technical report, 2025.

Clement Farabet and Olivier Lacombe. Gemma 4: Byte for byte, the most capable open models, 2026. URL https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/. Google DeepMind blog, accessed May 4, 2026.

Mohsen Fayyaz, Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Ryan A. Rossi, Trung Bui, Hinrich Schuetze, and Nanyun Peng. Steering MoE LLMs via expert (de)activation. In *The Fourteenth International Conference on Learning Representations*, 2026.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Hongcan Guo, Haolang Lu, Guoshun Nan, Bolun Chu, Jialin Zhuang, Yuan Yang, Wenhao Che, Xinye Cao, Sicong Leng, Qimei Cui, and Xudong Jiang. Advancing expert specialization for better MoE. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Jaehan Kim, Minkyoo Song, Seungwon Shin, and Sooel Son. SafeMoE: Safe fine-tuning for MoE LLMs by aligning harmful input routing. In *The Fourteenth International Conference on Learning Representations*, 2026.

ZhengLin Lai, Mengyao Liao, Bingzhe Wu, Dong Xu, Zebin Zhao, Zhihang Yuan, Chao Fan, and Jianqiang Li. SAFEx: Analyzing vulnerabilities of MoE-Based LLMs via stable safety-critical expert identification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Wengang Li, Lingqi Zhang, Toshio Endo, and Mohamed Wahib. Understanding cross-layer contributions to mixture-of-experts routing in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024.

Ang Lv, Jin Ma, Yiyuan Ma, and Siyuan Qiao. Coupling experts and routers in mixture-of-experts via an auxiliary loss. In *The Fourteenth International Conference on Learning Representations*, 2026.

María Luisa Menéndez, Julio Angel Pardo, Leandro Pardo, and María del C Pardo. The Jensen-Shannon Divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997.

Ashwinee Panda, Vatsal Baherwani, Zain Sarwar, Benjamin Thérien, Sambit Sahu, Tom Goldstein, and Supriyo Chakraborty. Dense backpropagation improves training for sparse mixture-of-experts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025.

GLM Team. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models, 2025.

Mengru Wang, Xingyu Chen, Yue Wang, Zhiwei He, Jiahao Xu, Tian Liang, Qiuzhi Liu, Yunzhi Yao, Wenxuan Wang, Ruotian Ma, Haitao Mi, Ningyu Zhang, Zhaopeng Tu, Xiaolong Li, and Dong Yu. Two experts are all you need for steering thinking: Reinforcing reasoning in MoE models without additional training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

A Experimental Setup

A.1 Model Specifications

We evaluate five MoE models from the Qwen (Yang et al., 2025), Gemma (Farabet & Lacombe, 2026), and GLM (Team, 2025) families. The models range from 26B to 235B total parameters and vary in expert count (64–256), top-K (4 or 8), and the presence of a shared expert. Table 5 summarizes their specifications.

Qwen3.5-35B-A3B, Gemma-4-26B-A4B-it, and GLM-4.7-Flash employ a shared expert that is activated for every token regardless of the router decision. All routing statistics reported in this paper cover only the *routed* experts, since the shared expert does not participate in the top-K selection.

A.2 Datasets

For GSM8K and HotpotQA we use fixed 500-question subsets, while MATH-500 and HumanEval use their full evaluation sets. The DAPO-Math subset is constructed by generating rollouts from Qwen3-30B-A3B with a 10,240-token budget and retaining 500 questions with both correct and incorrect outcomes. This initial screening differs from the subsequent 16-rollout evaluation and its eligibility checks. The outcome analyses use the recorded thinking spans, with answer extraction and correctness judgments following dataset-specific rules.

A.3 Outcome Analysis Coverage

Table 7 reports the mixed-question support for routing differences and linear occupancy probes across the 24 available model–dataset combinations. Selection and representation diagnostics use their own eligibility criteria, with sample counts reported alongside their results.

A.4 Infrastructure

Generation. The outcome-evaluation rollouts are generated with vLLM, sampling 16 responses per question at temperature 0.7 with thinking mode enabled. The maximum new-token budget is 10,240 for most datasets and 20,480 for DAPO-Math.

Routing capture. Expert assignments are recorded by replaying saved trajectories with temperature 0 and max_tokens $= 1$. The top-K expert indices at each MoE layer are logged for every token position. Results are stored in `.npz` format.

A.5 Statistical Procedures

Questions are the units of train–test splitting and uncertainty estimation for the outcome analyses. Method comparisons use paired resampling of the same questions, while each experiment retains its stated aggregation weights. RD uses minority-count weights, and the reported probe metrics average equally across eligible questions. Permutation testing and multiple-comparison correction are specified in Appendix B.2, with selection uncertainty in Appendix D.1. Confidence intervals describe variation across questions. The outcome, readout, and selection protocols retain their respective aggregation and eligibility rules.

Table 5: MoE model specifications used in our experiments.

Model	Params	Experts	Top-K	MoE layers	Shared expert	Hidden size
Qwen3-30B-A3B	30B	128	8	48	No	2048
Qwen3.5-35B-A3B	35B	256	8	40	Yes	2048
Qwen3-235B-A22B	235B	128	8	94	No	4096
Gemma-4-26B-A4B-it	26B	128	8	30	Yes	2816
GLM-4.7-Flash	30B	64	4	47	Yes	2048

Table 6: Datasets used in our experiments.

Dataset	Domain	Questions	Evaluation	Source
MATH-500	Math	500	Mathematical equivalence	Lightman et al. (2024)
GSM8K	Math	500	Numerical answer matching	Cobbe et al. (2021)
DAPO-Math	Math	500	Integer answer matching	Yu et al. (2025)
HotpotQA	QA	500	Normalized answer matching	Yang et al. (2018)
HumanEval	Code	164	Test execution	Chen et al. (2021)

Table 7: Mixed-question counts for routing differences and linear occupancy probes. A dash marks unavailable routing data. Small counts limit the conclusions supported by a setting.

Model	DAPO-Math	MATH-500	GSM8K	HotpotQA	HumanEval
Qwen3-30B-A3B	247	15	22	62	23
Qwen3.5-35B-A3B	32	–	7	48	10
Qwen3-235B-A22B	158	15	25	123	16
Gemma-4-26B-A4B-it	75	3	18	55	2
GLM-4.7-Flash	150	31	38	110	46

B Outcome Association

B.1 Routing Difference Algorithm

Algorithm 1 specifies the balanced estimator used in Section 3. Each rollout contributes a normalized occupancy distribution, and group means are formed from these distributions. We average signed differences over $S = 32$ balanced draws before taking their absolute values. The outcome subsets are $\mathcal{C}_q = \{r \in \mathcal{R}_q : c_r = 1\}$ and $\mathcal{I}_q = \{r \in \mathcal{R}_q : c_r = 0\}$, with sizes n_q^+ and n_q^-.

The group means in Eq. 2 are $\bar{p}^{\pm}_{q,l,\phi} = S^{-1} \sum_{s=1}^{S} \bar{p}^{\pm,s}_{q,l,\phi}$. For a fixed difference vector, the alternative expert-wise mean absolute summary equals $2\mathrm{RD}/E$. We use RD throughout the analysis to avoid this explicit division by expert count, while calibrating finite-sample separation with a matched null.

B.2 Matched Null, Significance, and Complete Results

Each of $P = 1000$ permutations shuffles correctness labels within each question while preserving the correct and incorrect counts. We then repeat balanced subsampling and the same layer, phase, and minority-weighted question aggregation. The one-sided permutation value is

$$p = \frac{1 + \sum_{b=1}^{P} \mathbf{1}\{\mathrm{RD}^{(b)} \geq \mathrm{RD}_{\mathrm{obs}}\}}{P + 1}, \tag{5}$$

and Holm correction is applied across the 24 available model–dataset combinations. With $P = 1000$ permutations, the smallest attainable one-sided value is $1/(P + 1) \approx 0.001$, which is the resolution of the reported permutation values. Question-bootstrap intervals resample questions with replacement and recompute the weighted mean. The null interval instead describes the central 95% of permutation statistics, and excess RD is reported as a point difference between the observation and the null mean. Table 8 reports all available combinations, including settings with limited question support.

Across the full matrix, 23 of 24 excess values are positive and 21 settings pass Holm correction. The passing counts are 5/5 on DAPO-Math, 3/5 on GSM8K, 5/5 on HotpotQA, 5/5 on HumanEval, and

Algorithm 1 Balanced Routing Difference Using Total Variation

Input Occupancies $p_{l,\phi}^{(r)}$ and labels c_r for $r \in \mathcal{R}_q$, phases Φ, layers $1, \ldots, L$, and $S = 32$
for each mixed question q **do**
  Form correct and incorrect subsets $\mathcal{C}_q$ and $\mathcal{I}_q$
  $m_q \leftarrow \min(n_q^+, n_q^-)$
  for $s = 1, \ldots, S$ **do**
    Sample m_q rollouts from each subset without replacement
    Compute their mean occupancies $\bar{p}_{q,l,\phi}^{+,s}$ and $\bar{p}_{q,l,\phi}^{-,s}$ for every layer and phase
    $\delta_{q,l,\phi}^{s} \leftarrow \bar{p}_{q,l,\phi}^{+,s} - \bar{p}_{q,l,\phi}^{-,s}$
  end for
  $\mathrm{RD}_{q,l,\phi} \leftarrow \frac{1}{2} \sum_{e=1}^{E} \left| \frac{1}{S} \sum_{s=1}^{S} \delta_{q,l,\phi,e}^{s} \right|$
end for
Return $\mathrm{RD} = \frac{1}{L|\Phi|} \sum_{l,\phi} \frac{\sum_q m_q \mathrm{RD}_{q,l,\phi}}{\sum_q m_q}$

Table 8: **Complete null-calibrated routing differences.** RD is total variation distance. Observed intervals are 95% question-bootstrap confidence intervals, while null intervals cover the central 95% of permutations. Excess is observed minus mean null RD, and p_{adj} uses Holm correction across the 24 available model–dataset combinations. Small-question settings are descriptive.

Model	Dataset	Q	RD [95% CI]	Null [95% interval]	Excess	p_{adj}
Qwen3-30B-A3B	DAPO-Math	247	0.0657 [0.0611, 0.0704]	0.0496 [0.0478, 0.0517]	+0.0161	0.02398
Qwen3.5-35B-A3B	DAPO-Math	32	0.1062 [0.0981, 0.1170]	0.0858 [0.0798, 0.0931]	+0.0204	0.02398
Qwen3-235B-A22B	DAPO-Math	158	0.0657 [0.0598, 0.0715]	0.0519 [0.0497, 0.0543]	+0.0138	0.02398
Gemma-4-26B-A4B-it	DAPO-Math	75	0.0466 [0.0421, 0.0520]	0.0405 [0.0383, 0.0427]	+0.0061	0.02398
GLM-4.7-Flash	DAPO-Math	150	0.0793 [0.0729, 0.0863]	0.0681 [0.0655, 0.0709]	+0.0112	0.02398
Qwen3-30B-A3B	MATH-500	15	0.1018 [0.0755, 0.1384]	0.0640 [0.0556, 0.0733]	+0.0378	0.02398
Qwen3-235B-A22B	MATH-500	15	0.1038 [0.0792, 0.1349]	0.0735 [0.0652, 0.0829]	+0.0303	0.02398
Gemma-4-26B-A4B-it	MATH-500	3	0.0519 [0.0451, 0.0607]	0.0534 [0.0427, 0.0663]	−0.0016	0.57542
GLM-4.7-Flash	MATH-500	31	0.1228 [0.1061, 0.1442]	0.1037 [0.0964, 0.1123]	+0.0191	0.02398
Qwen3-30B-A3B	GSM8K	22	0.0737 [0.0558, 0.1013]	0.0554 [0.0497, 0.0618]	+0.0184	0.02398
Qwen3.5-35B-A3B	GSM8K	7	0.1028 [0.0906, 0.1332]	0.0935 [0.0767, 0.1151]	+0.0093	0.36164
Qwen3-235B-A22B	GSM8K	25	0.0828 [0.0570, 0.1133]	0.0558 [0.0481, 0.0663]	+0.0271	0.02398
Gemma-4-26B-A4B-it	GSM8K	18	0.0460 [0.0390, 0.0566]	0.0392 [0.0338, 0.0468]	+0.0068	0.11389
GLM-4.7-Flash	GSM8K	38	0.1256 [0.1065, 0.1482]	0.0962 [0.0895, 0.1033]	+0.0294	0.02398
Qwen3-30B-A3B	HotpotQA	62	0.0833 [0.0745, 0.0940]	0.0628 [0.0591, 0.0670]	+0.0205	0.02398
Qwen3.5-35B-A3B	HotpotQA	48	0.0939 [0.0849, 0.1055]	0.0771 [0.0726, 0.0818]	+0.0168	0.02398
Qwen3-235B-A22B	HotpotQA	123	0.1066 [0.0993, 0.1163]	0.0907 [0.0872, 0.0945]	+0.0159	0.02398
Gemma-4-26B-A4B-it	HotpotQA	55	0.0615 [0.0541, 0.0696]	0.0503 [0.0475, 0.0535]	+0.0112	0.02398
GLM-4.7-Flash	HotpotQA	110	0.0960 [0.0906, 0.1017]	0.0807 [0.0778, 0.0840]	+0.0153	0.02398
Qwen3-30B-A3B	HumanEval	23	0.0761 [0.0635, 0.0917]	0.0618 [0.0559, 0.0682]	+0.0144	0.02398
Qwen3.5-35B-A3B	HumanEval	10	0.1316 [0.0963, 0.1650]	0.1138 [0.1014, 0.1263]	+0.0178	0.02398
Qwen3-235B-A22B	HumanEval	16	0.0899 [0.0780, 0.1048]	0.0712 [0.0643, 0.0792]	+0.0187	0.02398
Gemma-4-26B-A4B-it	HumanEval	2	0.0690 [0.0571, 0.1527]	0.0497 [0.0372, 0.0671]	+0.0193	0.04795
GLM-4.7-Flash	HumanEval	46	0.1394 [0.1248, 0.1586]	0.1219 [0.1144, 0.1308]	+0.0174	0.02398

3/4 on MATH-500. This broad pattern connects the representative rows in Table 1 to all available settings.

B.3 Length Control and Same-Label References

As complements to the permutation null, Table 9 reports two diagnostics on the same RD pipeline. The length-matched estimate subsamples the correct and incorrect groups so that they overlap in rollout length before computing RD, on the questions where this matching is feasible. Same-label references provide a sampling floor: C/C and I/I split the correct-only and incorrect-only rollouts of a

question into random halves and compute the same TV statistic between the halves. They describe same-label variation on different support sets and cannot replace the matched null, which preserves each question's group sizes and candidate pool.

Length-matched RD remains at a similar scale across the DAPO comparisons (Table 9 and Figure 2). For Qwen3-30B, it changes from 0.0657 to 0.0703 after matching. Same-label comparisons also yield positive distances, illustrating the finite-sample variation calibrated by the matched permutation null.

B.4 Progress-Bin Sensitivity

The primary outcome representation divides the thinking span into paragraph-based segments and groups them by relative segment position. The resulting early, middle, and late bins describe generation progress and need not contain equal token counts. They do not assign semantic functions to the bins. A sensitivity comparison replaces these boundaries with exact token thirds while retaining the same DAPO-Math rollouts, correctness labels, question folds, pair examples, and length baseline.

Replacing paragraph bins with equal-token thirds produces small changes in the three available comparisons (Table 10), with the AUC comparison visualized in Figure 3. The sensitivity RD decreases by 0.0017 to 0.0029, while routing AUC changes by -2.33, $+0.44$, and -0.10 percentage points. The observed readout pattern is broadly similar across these two ways of representing generation progress.

Table 9: **Length control and same-label sampling references.** Length-matched RD recomputes RD after matching group lengths on the supported questions, e.g., 245 of 247 for Qwen3-30B-A3B on DAPO-Math. C/C and I/I report RD between random half-splits of correct-only and incorrect-only rollouts. Parentheses give supporting question counts, which differ across columns.

Model	Dataset	Length-matched RD (Q)	C/C RD (Q)	I/I RD (Q)
Qwen3-30B-A3B	DAPO-Math	0.0703 (245)	0.0530 (225)	0.0634 (149)
Qwen3.5-35B-A3B	DAPO-Math	0.1210 (30)	0.0751 (14)	0.0838 (17)
Gemma-4-26B-A4B-it	DAPO-Math	0.0540 (71)	0.0429 (64)	0.0444 (39)
GLM-4.7-Flash	DAPO-Math	0.0852 (137)	0.0646 (121)	0.0740 (69)
Qwen3-235B-A22B	DAPO-Math	0.0713 (152)	0.0531 (142)	0.0615 (87)
Qwen3-30B-A3B	MATH-500	0.0988 (14)	0.0567 (12)	0.0539 (5)
Qwen3-30B-A3B	GSM8K	0.0530 (15)	0.0567 (18)	0.0507 (11)
Qwen3-30B-A3B	HotpotQA	0.0813 (58)	0.0654 (43)	0.0723 (48)
Qwen3-30B-A3B	HumanEval	0.0852 (22)	0.0659 (20)	0.0797 (12)

Table 10: **Sensitivity to progress-bin boundaries.** Equal-token thirds replace paragraph-based bins on the same DAPO-Math rollouts, labels, question folds, pair examples, and length baseline. RD columns compare the sensitivity pipeline's RD estimates under the two segmentations, and AUC columns compare the linear routing probe. Differences are equal-token minus paragraph values, with paired 95% question-bootstrap intervals.

Model	Q	RD par.	RD equal	ΔRD [95% CI]	AUC par.	AUC equal	ΔAUC [95% CI]
Gemma-4-26B-A4B-it	75	0.0575	0.0546	$-0.0029\,[-0.0056, -0.0003]$	0.5000	0.4767	$-0.0233\,[-0.0917, 0.0517]$
Qwen3-235B-A22B	158	0.0750	0.0733	$-0.0017\,[-0.0030, -0.0006]$	0.5862	0.5906	$+0.0044\,[-0.0324, 0.0423]$
Qwen3-30B-A3B	247	0.0761	0.0742	$-0.0020\,[-0.0029, -0.0011]$	0.5774	0.5764	$-0.0010\,[-0.0266, 0.0266]$

C Readout Results

C.1 Pair-order AUC and Probe Configuration

The linear probe uses L_2-regularized logistic regression, while the nonlinear sensitivity uses a one-layer MLP with 64 hidden units after training-fold dimensionality reduction to at most 64 SVD components. Both use the within-question feature differences $x_{r_a} - x_{r_b}$ from Section 4, with both orientations included. We compute pair-order AUC within each test question and then take an unweighted mean across questions. The reported differences compare routing-only and length-only readouts under the same splits, with paired question-bootstrap confidence intervals. Tables 11 and 12 report all available settings, with $Q < 30$ treated as descriptive. The nonlinear probe is unavailable for Gemma-4-26B-A4B-it on HumanEval, where only two mixed questions are available.

The nonlinear sensitivity varies across settings rather than uniformly strengthening the linear readout. On DAPO-Math, Qwen3.5 and Gemma improve over their MLP length baselines, while GLM and the two Qwen3 models have lower point estimates. This variation emphasizes that the measured quality signal depends on the representation and readout used, motivating evaluation through the actual selection task.

Table 11: **Complete linear occupancy-probe results.** AUC is averaged over questions using pair-order examples, on a zero-to-one scale. Differences are routing-only minus length-only on the same scale, with paired 95% question-bootstrap confidence intervals. Settings with $Q < 30$ are descriptive, and a dash marks an unavailable result.

Model	Dataset	Q	Length AUC	Routing AUC	Difference [95% CI]
Qwen3-30B-A3B	DAPO-Math	247	0.5736	0.5774	$+0.0038\,[-0.0564, 0.0663]$
Qwen3.5-35B-A3B	DAPO-Math	32	0.3791	0.5703	$+0.1912\,[0.0058, 0.3921]$
Qwen3-235B-A22B	DAPO-Math	158	0.5720	0.5862	$+0.0142\,[-0.0538, 0.0867]$
Gemma-4-26B-A4B-it	DAPO-Math	75	0.4092	0.5000	$+0.0908\,[-0.0175, 0.1908]$
GLM-4.7-Flash	DAPO-Math	150	0.5033	0.5950	$+0.0917\,[0.0133, 0.1754]$
Qwen3-30B-A3B	MATH-500	15	0.4667	0.5000	$+0.0333\,[-0.2210, 0.2959]$
Qwen3-235B-A22B	MATH-500	15	0.5208	0.2333	$-0.2875\,[-0.4960, -0.0708]$
Gemma-4-26B-A4B-it	MATH-500	3	0.2708	0.9167	$+0.6458\,[0.3750, 0.9375]$
GLM-4.7-Flash	MATH-500	31	0.5343	0.7258	$+0.1915\,[0.0504, 0.3427]$
Qwen3-30B-A3B	GSM8K	22	0.7102	0.5682	$-0.1420\,[-0.3125, 0.0227]$
Qwen3.5-35B-A3B	GSM8K	7	0.6875	0.5446	$-0.1429\,[-0.5002, 0.1969]$
Qwen3-235B-A22B	GSM8K	25	0.7050	0.7375	$+0.0325\,[-0.0850, 0.1400]$
Gemma-4-26B-A4B-it	GSM8K	18	0.3785	0.4826	$+0.1042\,[-0.1111, 0.3022]$
GLM-4.7-Flash	GSM8K	38	0.7286	0.7451	$+0.0164\,[-0.0921, 0.1168]$
Qwen3-30B-A3B	HotpotQA	62	0.4385	0.5726	$+0.1341\,[-0.0192, 0.2944]$
Qwen3.5-35B-A3B	HotpotQA	48	0.5026	0.4609	$-0.0417\,[-0.1875, 0.1003]$
Qwen3-235B-A22B	HotpotQA	123	0.5567	0.5396	$-0.0170\,[-0.1077, 0.0668]$
Gemma-4-26B-A4B-it	HotpotQA	55	0.4756	0.5114	$+0.0358\,[-0.0955, 0.1569]$
GLM-4.7-Flash	HotpotQA	110	0.5023	0.5114	$+0.0091\,[-0.0767, 0.0972]$
Qwen3-30B-A3B	HumanEval	23	0.5272	0.6685	$+0.1413\,[-0.0571, 0.3451]$
Qwen3.5-35B-A3B	HumanEval	10	0.8938	0.7812	$-0.1125\,[-0.3250, 0.0563]$
Qwen3-235B-A22B	HumanEval	16	0.6250	0.5742	$-0.0508\,[-0.2227, 0.1133]$
Gemma-4-26B-A4B-it	HumanEval	2	0.7812	0.8438	$+0.0625\,[0.0625, 0.0625]$
GLM-4.7-Flash	HumanEval	46	0.4579	0.6155	$+0.1576\,[-0.0089, 0.2956]$

Table 12: **Complete nonlinear occupancy-probe results.** AUC is averaged over questions using pair-order examples, on a zero-to-one scale. Differences are routing-only minus length-only on the same scale, with paired 95% question-bootstrap confidence intervals. Settings with $Q < 30$ are descriptive, and a dash marks an unavailable result.

Model	Dataset	Q	Length AUC	Routing AUC	Difference [95% CI]
Qwen3-30B-A3B	DAPO-Math	247	0.5701	0.5676	$-0.0025\,[-0.0683, 0.0600]$
Qwen3.5-35B-A3B	DAPO-Math	32	0.3772	0.5358	$+0.1586\,[-0.0432, 0.3746]$
Qwen3-235B-A22B	DAPO-Math	158	0.5720	0.5680	$-0.0040\,[-0.0791, 0.0811]$
Gemma-4-26B-A4B-it	DAPO-Math	75	0.4508	0.5433	$+0.0925\,[-0.0267, 0.2126]$
GLM-4.7-Flash	DAPO-Math	150	0.4892	0.4629	$-0.0262\,[-0.1054, 0.0550]$
Qwen3-30B-A3B	MATH-500	15	0.5083	0.5852	$+0.0769\,[-0.2214, 0.3519]$
Qwen3-235B-A22B	MATH-500	15	0.6958	0.4875	$-0.2083\,[-0.4626, 0.0459]$
Gemma-4-26B-A4B-it	MATH-500	3	0.2708	0.8125	$+0.5417\,[0.0000, 1.0000]$
GLM-4.7-Flash	MATH-500	31	0.5096	0.5927	$+0.0831\,[-0.0600, 0.2247]$
Qwen3-30B-A3B	GSM8K	22	0.7045	0.5426	$-0.1619\,[-0.3580, 0.0398]$
Qwen3.5-35B-A3B	GSM8K	7	0.6875	0.6429	$-0.0446\,[-0.3393, 0.2594]$
Qwen3-235B-A22B	GSM8K	25	0.7050	0.5450	$-0.1600\,[-0.3401, 0.0326]$
Gemma-4-26B-A4B-it	GSM8K	18	0.2604	0.5590	$+0.2986\,[0.0590, 0.4897]$
GLM-4.7-Flash	GSM8K	38	0.7286	0.6743	$-0.0543\,[-0.1809, 0.0823]$
Qwen3-30B-A3B	HotpotQA	62	0.4536	0.5665	$+0.1129\,[-0.0182, 0.2561]$
Qwen3.5-35B-A3B	HotpotQA	48	0.5117	0.5000	$-0.0117\,[-0.1576, 0.1263]$
Qwen3-235B-A22B	HotpotQA	123	0.5521	0.5513	$-0.0008\,[-0.0803, 0.0757]$
Gemma-4-26B-A4B-it	HotpotQA	55	0.4369	0.4091	$-0.0278\,[-0.1324, 0.0842]$
GLM-4.7-Flash	HotpotQA	110	0.5125	0.5080	$-0.0045\,[-0.1011, 0.0909]$
Qwen3-30B-A3B	HumanEval	23	0.3777	0.4946	$+0.1168\,[-0.1087, 0.3316]$
Qwen3.5-35B-A3B	HumanEval	10	0.8938	0.7063	$-0.1875\,[-0.3875, 0.0375]$
Qwen3-235B-A22B	HumanEval	16	0.4961	0.3945	$-0.1016\,[-0.3361, 0.1485]$
Gemma-4-26B-A4B-it	HumanEval	2	–	–	–
GLM-4.7-Flash	HumanEval	46	0.5503	0.5014	$-0.0489\,[-0.1889, 0.0877]$

D CANDIDATE SELECTION

D.1 CANDIDATE SCORING PROTOCOL AND EVALUATION POOLS

The learned-selection comparison covers five models on DAPO-Math, the same five on HotpotQA, and GLM on HumanEval. Table 13 gives the evaluation populations. Every reported evaluation question contains both correct and incorrect candidates with judgeable labels and valid saved features. All candidates from a question stay in the same outer fold, and candidate scores are evaluated on held-out questions. We fit the scorers using primal LIBLINEAR with $C = 1$, tolerance 10^{-4}, and at most 5,000 iterations.

Gemma on DAPO has 80 retained training/evaluation questions in the saved run, of which 75 are mixed. All tables and budget curves in this paper evaluate those 75 mixed questions, while the original training folds, including five non-mixed questions, are retained. All six HotpotQA/HumanEval settings are naturally mixed, so their retained and evaluated question sets coincide.

Top-1 checks the highest-scoring candidate, and Coverage@2 checks whether either of the two highest-scoring candidates is correct. Ties are broken by score descending and then saved candidate index ascending. Point estimates are computed from the saved test-fold predictions. The candidate-ranking AUC and the Top-1 and Coverage@2 increments use paired question bootstrap with 10,000 resamples and seed 20260920. The comparisons are exploratory and do not apply multiplicity correction.

D.2 CANDIDATE SCORER FEATURES AND RANKING AUC

S_B uses the recorded token count through $\log(1 + T)$ and four simple text statistics, consisting of log character count, log token count over the final 1,024 tokens, newline frequency, and digit frequency. S_{BR} appends the routing feature vector x_r from Section 2.2. Features are standardized using training-fold statistics. Both scorers use the same five outer question folds and the archived inner question folds, with three inner folds generated when an archived assignment is unavailable. Inner out-of-fold scores provide training inputs for the pair models, and the outer model scores the held-out questions. The random seed is 20260920.

Candidate-ranking AUC is computed within each test question and averaged equally across questions (Table 15). Seven of the eleven settings show a positive increment. GLM on DAPO gains 7.0 percentage points, while GLM on HumanEval loses 12.1 points. These results accompany the retention outcomes and show that the additional ranking value of occupancy depends on the task and the existing features.

D.3 PAIR SELECTION PROTOCOL AND COMPLETE RESULTS

All unordered candidate pairs form the within-question pair set. A pair is labeled by *joint failure* when both candidates are incorrect. The pair models use the same L_2-regularized logistic configuration as the candidate scorers. P_0 summarizes the two base-score probabilities through their minimum, maximum, and absolute difference, combines log lengths through their mean and absolute difference, and includes text 3-gram Jaccard distance over the final 1,024 tokens. P_1 adds the same three summaries of the routing-augmented scores. P_2 further adds mean Jensen–Shannon distance between occupancy distributions over the recorded layer–phase slots, with the square root taken before averaging. The selected pair has the lowest predicted joint-failure probability, following Equation 4.

Table 16 reports coverage and the distance increment for every setting. Paired intervals for $P_2 - P_1$ use 1,000 question-bootstrap resamples with seed 20260920, retaining the original run protocol and recomputing Gemma-DAPO on its mixed subset. The increments vary in sign on HotpotQA and are mostly zero or negative on DAPO, with no individual positive interval excluding zero. Thus the distance feature does not supply a consistent additional retention benefit once scores and text information are available.

Tables 17 and 18 report the full prediction metrics. Log loss and Brier score average within questions and then across questions. Pair AUC uses questions with both joint-failure classes, whose counts are reported separately. Questions with a single pair-label class remain in coverage, log loss, and Brier evaluation. All Gemma-DAPO metrics shown here use the mixed evaluation subset.

Table 13: **Learned-selection evaluation pools.** Counts refer to mixed questions and their retained candidates and unordered pairs. Training folds are preserved from the converged runs.

Model	Dataset	Questions	Candidates	Pairs
Qwen3-30B	DAPO-Math	247	3893	28797
Qwen3.5	DAPO-Math	32	317	1695
Gemma	DAPO-Math	75	1031	6915
GLM	DAPO-Math	150	2231	15990
Qwen3-235B	DAPO-Math	158	2508	18678
Qwen3-30B	HotpotQA	62	975	7210
Qwen3.5	HotpotQA	48	717	5046
Gemma	HotpotQA	55	853	6255
GLM	HotpotQA	110	1718	12632
Qwen3-235B	HotpotQA	123	1873	13461
GLM	HumanEval	46	669	4676

Table 14: **Selection increments across all eleven settings.** Paired differences are $S_{BR} - S_B$, in percentage points, with 95% question-bootstrap intervals from 10,000 resamples. Absolute scores appear in Table 3.

Model	Dataset	Q	ΔTop-1 [95% CI]	ΔCoverage@2 [95% CI]
Qwen3-30B	DAPO-Math	247	$+3.64\,[-2.43, +9.72]$	$+2.43\,[-3.24, +8.10]$
Qwen3.5	DAPO-Math	32	$+12.50\,[-9.38, +34.38]$	$+9.38\,[-9.38, +28.12]$
Gemma	DAPO-Math	75	$-1.33\,[-14.67, +12.00]$	$-1.33\,[-13.33, +10.67]$
GLM	DAPO-Math	150	$+8.67\,[+0.00, +18.00]$	$+7.33\,[+0.67, +14.00]$
Qwen3-235B	DAPO-Math	158	$+1.27\,[-6.96, +9.49]$	$+0.63\,[-5.70, +6.96]$
Qwen3-30B	HotpotQA	62	$+1.61\,[-14.52, +17.74]$	$+9.68\,[-3.23, +22.58]$
Qwen3.5	HotpotQA	48	$+0.00\,[-14.58, +14.58]$	$-8.33\,[-22.92, +6.25]$
Gemma	HotpotQA	55	$-14.55\,[-27.27, -3.64]$	$-5.45\,[-16.36, +5.45]$
GLM	HotpotQA	110	$+9.09\,[+0.00, +18.18]$	$+1.82\,[-6.36, +10.00]$
Qwen3-235B	HotpotQA	123	$-2.44\,[-12.20, +7.32]$	$+1.63\,[-7.32, +10.57]$
GLM	HumanEval	46	$-2.17\,[-15.22, +10.87]$	$-4.35\,[-10.87, +0.00]$

Table 15: **Candidate-ranking AUC across all learned-selection settings.** AUC is computed within each question and averaged equally across questions, in percent. Differences are $S_{BR} - S_B$ in percentage points, with paired 95% intervals from 10,000 question-bootstrap resamples.

Model	Dataset	Q	S_B	S_{BR}	Difference [95% CI]
Qwen3-30B	DAPO-Math	247	56.17	56.49	$+0.33\,[-3.68, +4.32]$
Qwen3.5	DAPO-Math	32	51.90	57.04	$+5.15\,[-12.66, +22.23]$
Gemma	DAPO-Math	75	49.96	49.34	$-0.62\,[-9.46, +8.17]$
GLM	DAPO-Math	150	50.42	57.39	$+6.97\,[+0.90, +13.04]$
Qwen3-235B	DAPO-Math	158	53.19	56.33	$+3.14\,[-2.00, +8.25]$
Qwen3-30B	HotpotQA	62	49.71	54.86	$+5.15\,[-6.36, +16.30]$
Qwen3.5	HotpotQA	48	51.96	54.23	$+2.27\,[-8.53, +13.17]$
Gemma	HotpotQA	55	51.59	47.35	$-4.24\,[-12.14, +3.62]$
GLM	HotpotQA	110	50.93	56.04	$+5.11\,[-1.51, +11.91]$
Qwen3-235B	HotpotQA	123	55.92	52.77	$-3.15\,[-9.36, +2.94]$
GLM	HumanEval	46	68.62	56.48	$-12.14\,[-22.78, -1.20]$

D.4 Pooled DAPO Estimates

We summarize the five DAPO models by averaging their question-level paired differences. The equal-weight estimate gives each model weight $1/5$, and the question-weighted estimate gives model m weight $Q_m / \sum_j Q_j$. A stratified bootstrap resamples questions independently within each model and recomputes both estimates for 10,000 draws with seed 20260920. These summaries cover the 662 DAPO evaluation questions only. Table 19 uses the full-precision estimates before rounding.

Table 16: **Complete pair-selection coverage.** Coverage@2 is in percent. P_0 uses base scores, lengths and text distance, P_1 adds routing scores, and P_2 adds mean routing distance. Differences are percentage points with 1,000-resample paired 95% intervals.

Model	Dataset	Q	P_0	P_1	P_2	$P_2 - P_1$ [95% CI]
Qwen3-30B	DAPO-Math	247	92.31	91.09	91.50	$+0.40\,[-0.81, +2.02]$
Qwen3.5	DAPO-Math	32	81.25	78.12	71.88	$-6.25\,[-15.62, +0.00]$
Gemma	DAPO-Math	75	84.00	85.33	84.00	$-1.33\,[-4.00, +0.00]$
GLM	DAPO-Math	150	84.00	84.67	84.67	$+0.00\,[-4.00, +4.67]$
Qwen3-235B	DAPO-Math	158	87.34	86.71	83.54	$-3.16\,[-6.96, +0.63]$
Qwen3-30B	HotpotQA	62	82.26	80.65	80.65	$+0.00\,[+0.00, +0.00]$
Qwen3.5	HotpotQA	48	77.08	75.00	79.17	$+4.17\,[-4.17, +12.50]$
Gemma	HotpotQA	55	69.09	78.18	74.55	$-3.64\,[-14.55, +5.45]$
GLM	HotpotQA	110	80.00	79.09	83.64	$+4.55\,[+0.00, +10.00]$
Qwen3-235B	HotpotQA	123	82.11	81.30	82.93	$+1.63\,[-2.44, +5.69]$
GLM	HumanEval	46	97.83	97.83	97.83	$+0.00\,[+0.00, +0.00]$

Table 17: **Pair prediction metrics on DAPO-Math.** Log loss and Brier score are averaged within questions and then across questions. Pair AUC in percent uses only questions with both joint-failure classes, counted in Q_{AUC}.

Model	Dataset	Predictor	Log loss	Brier	AUC	Q_{AUC}
Qwen3-30B	DAPO-Math	P_0	0.4420	0.1375	61.20	202
Qwen3-30B	DAPO-Math	P_1	0.4391	0.1367	60.85	202
Qwen3-30B	DAPO-Math	P_2	0.4401	0.1370	61.03	202
Qwen3.5	DAPO-Math	P_0	0.6452	0.2251	55.24	23
Qwen3.5	DAPO-Math	P_1	0.5911	0.1999	56.48	23
Qwen3.5	DAPO-Math	P_2	0.6076	0.2059	54.51	23
Gemma	DAPO-Math	P_0	0.5533	0.1815	64.74	55
Gemma	DAPO-Math	P_1	0.5631	0.1849	63.16	55
Gemma	DAPO-Math	P_2	0.5616	0.1846	62.67	55
GLM	DAPO-Math	P_0	0.5005	0.1593	51.72	97
GLM	DAPO-Math	P_1	0.4953	0.1577	56.30	97
GLM	DAPO-Math	P_2	0.4972	0.1583	56.99	97
Qwen3-235B	DAPO-Math	P_0	0.4777	0.1502	56.03	118
Qwen3-235B	DAPO-Math	P_1	0.4805	0.1511	55.95	118
Qwen3-235B	DAPO-Math	P_2	0.4824	0.1516	56.87	118

Both weighting schemes give positive learning-score increments and negative distance increments. The learning-score intervals include zero, while the equal-weight distance interval is wholly negative. Leaving out each DAPO model in turn preserves the positive equal-weight estimates for Top-1 (3.06 to 6.52 points), Coverage@2 (2.27 to 4.94), and candidate AUC (2.00 to 3.90). The distance increment remains negative in all five leave-one-model-out estimates, ranging from -2.69 to -1.02 points. The question-weighted Top-1 lower endpoint is exactly zero under the discrete bootstrap.

For an additional descriptive combination, we divide each model's paired mean by its bootstrap standard error and combine the resulting values using equal-weight Stouffer aggregation. The two-sided values are 0.061 for Top-1, 0.072 for Coverage@2, 0.073 for candidate AUC, and 0.109 for $P_2 - P_1$. These summaries describe the overall direction alongside the setting-specific results.

D.5 Selection Budgets

At each budget $k \in \{1, \ldots, 6\}$, we rank candidates by their saved scores and evaluate only mixed questions with at least k candidates. The evaluated population can therefore change with k. At $k = 2$, both the DAPO and cross-task curves agree with Table 3. The random-selection expectation for a pool with n candidates and w incorrect candidates is $1 - \binom{w}{k}/\binom{n}{k}$, using zero for $\binom{w}{k}$ when $w < k$.

The curves show how the same scoring comparison changes with the retention budget. On DAPO, GLM gains at all six budgets, Qwen3.5 gains at budgets one through four and ties thereafter, and

Table 18: **Pair prediction metrics on HotpotQA and HumanEval.** Log loss and Brier score are averaged within questions and then across questions. Pair AUC in percent uses only questions with both joint-failure classes, counted in Q_{AUC}.

Model	Dataset	Predictor	Log loss	Brier	AUC	Q_{AUC}
Qwen3-30B	HotpotQA	P_0	0.6354	0.2205	73.67	54
Qwen3-30B	HotpotQA	P_1	0.6445	0.2243	73.03	54
Qwen3-30B	HotpotQA	P_2	0.6460	0.2249	72.88	54
Qwen3.5	HotpotQA	P_0	0.5950	0.2025	51.21	42
Qwen3.5	HotpotQA	P_1	0.6156	0.2110	50.33	42
Qwen3.5	HotpotQA	P_2	0.6265	0.2155	50.22	42
Gemma	HotpotQA	P_0	0.6616	0.2315	51.40	46
Gemma	HotpotQA	P_1	0.6706	0.2352	51.52	46
Gemma	HotpotQA	P_2	0.6747	0.2366	52.72	46
GLM	HotpotQA	P_0	0.5688	0.1911	63.77	92
GLM	HotpotQA	P_1	0.5632	0.1886	65.13	92
GLM	HotpotQA	P_2	0.5629	0.1883	64.79	92
Qwen3-235B	HotpotQA	P_0	0.5892	0.1986	61.35	92
Qwen3-235B	HotpotQA	P_1	0.5846	0.1967	61.01	92
Qwen3-235B	HotpotQA	P_2	0.5835	0.1966	61.74	92
GLM	HumanEval	P_0	0.1872	0.0446	62.90	22
GLM	HumanEval	P_1	0.1875	0.0446	59.55	22
GLM	HumanEval	P_2	0.1886	0.0447	60.47	22

Table 19: **Pooled DAPO increments.** Differences and 95% intervals are percentage points. The first three rows compare $S_{BR} - S_B$, and the final row isolates mean routing distance in pair selection. These exploratory intervals are not multiplicity corrected.

Metric	Equal model weight	Question weight
Top-1	$+4.95\,[-0.93, +10.80]$	$+4.08\,[0.00, +8.16]$
Coverage@2	$+3.69\,[-1.43, +8.75]$	$+3.02\,[-0.30, +6.34]$
Candidate-ranking AUC	$+2.99\,[-1.27, +7.32]$	$+2.63\,[-0.08, +5.40]$
$P_2 - P_1$ Coverage@2	$-2.07\,[-4.42, -0.13]$	$-1.06\,[-2.57, +0.45]$

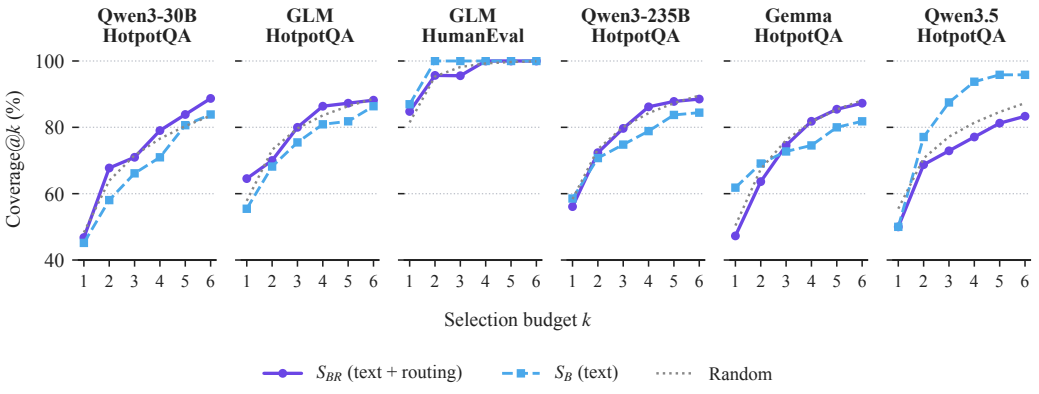


Figure 7: **Cross-task scoring across selection budgets.** Solid curves show S_{BR}, dashed curves show S_B, and dotted curves show the random-selection expectation. Each budget uses mixed questions with at least k candidates. The six panels extend the five DAPO comparisons in Figure 5.

the other models show smaller or changing increments. The cross-task curves retain all six settings, including the decreases on HumanEval and on Gemma and Qwen3.5 at budget two on HotpotQA.

D.6 Diversity Selection from the Eligible Pool

Candidate pools. We reuse the archived selection pools built from judgeable, non-truncated rollouts with valid routing captures. Each pool contains both correct and incorrect candidates, and the reported comparisons require at least six candidates before applying any selector. The support therefore differs from the routing-difference and probe analyses. Rare-correct questions contain exactly one correct candidate in this full eligible pool, before score shortlisting.

Diversity selectors. Routing distance is the mean Jensen–Shannon distance across the layer–phase occupancy distributions stored in each routing sketch. The implementation takes the square root of Jensen–Shannon divergence using natural logarithms. Text distance is one minus the Jaccard similarity of token 3-gram sets. Each diversity selector starts from the most distant pair and repeatedly adds the candidate with the largest minimum distance to the selected set. Ties favor shorter rollouts and then lower rollout indices. Random selection samples k candidates uniformly without replacement. All three selectors use the full eligible pool and do not use the synthetic scores.

Table 20 summarizes the pure-diversity comparisons. Routing-diverse selection improves over random sampling in some settings, with a median Coverage@4 gain of 2.47 percentage points, but it shows no reliable advantage over text diversity, with a median difference of -0.32 points and no interval entirely above zero. These results use the full eligible pools without scores and complement the pair-model comparison in §5.2.

D.7 Synthetic-Score Controlled Comparison

For candidate r with correctness label c_r, the noisy-oracle provider generates $s_r = c_r + \epsilon_r$, where $\epsilon_r \sim \mathcal{N}(0, \sigma^2)$. The two score-based selectors share the same score realizations and their top-six shortlist. Score-only returns the highest-scoring k candidates. Score+routing starts with the highest-scoring candidate and adds candidates that maximize the minimum routing distance to those already selected. Ties favor higher scores, then shorter rollouts and lower indices. The shared synthetic scores hold the scoring condition fixed while the selection rule changes.

Tables 21 and 22 report selector metrics and paired differences for all 24 available settings. The displayed metric means and differences are rounded independently. Coverage@4 is close to one for many score-only baselines. Rare-retain@2 complements this view by focusing on whether the only correct candidate survives a smaller budget.

D.8 Rare-Retention Results

Rare-retain@2 measures retention when the original eligible pool contains exactly one correct candidate and only two can be kept. Among the nine settings with at least five rare-correct questions, all Rare-retain@2 point differences are negative and six intervals exclude zero. For Qwen3-30B-A3B on DAPO-Math, retention falls from 78.86% to 73.43%, giving a difference of -5.43 points with a 95% interval of $[-10.29, -0.71]$. This consistent decrease shows that the tested diversity rule can remove the sole correct answer from the retained set.

D.9 Selection Metrics and Uncertainty

Coverage@4 is the fraction of eligible questions for which the selected set contains at least one correct candidate. Rare-retain@2 is the corresponding retention rate among questions with exactly one correct candidate, using a selection budget of two. Neither metric evaluates a procedure that chooses one final answer from the retained set.

We average score noise and random-selection draws within each question before estimating population means. Each paired comparison uses the same questions and subtracts their two selection metrics before question-level bootstrap. Intervals are computed separately for each setting and comparison without a cross-setting multiplicity adjustment. Counts of positive or negative intervals describe these individual comparisons, and the median difference across settings is not a pooled question-level effect. Settings with fewer than five rare-correct questions are retained as descriptive results. With very few questions, a bootstrap interval can be degenerate despite substantial uncertainty about generalization.

Table 20: **Coverage@4 diversity-selection summary.** Median paired differences across 24 settings are in percentage points. Positive and negative columns count individual 95% question-bootstrap intervals lying wholly above or below zero.

Comparison	Median difference	Positive CI	Negative CI
Routing-Diverse − Random	$+2.47$	4/24	0/24
Routing-Diverse − Text-Diverse	-0.32	0/24	2/24

Table 21: **Complete Coverage@4 under synthetic scores.** S denotes Score-only and S+R denotes Score+routing with a shortlist of six. Metrics are percentages, and differences are percentage points. Brackets give 95% question-bootstrap intervals, paired for differences. Q counts eligible mixed questions.

Model	Dataset	Q	S [95% CI]	S+R [95% CI]	Difference [95% CI]
Qwen3-30B-A3B	DAPO-Math	247	99.68 [99.44, 99.88]	99.57 [99.22, 99.84]	−0.12 [−0.30, 0.03]
Qwen3.5-35B-A3B	DAPO-Math	26	98.19 [96.73, 99.50]	98.65 [97.42, 99.62]	+0.46 [−0.77, 1.58]
Qwen3-235B-A22B	DAPO-Math	158	99.29 [98.87, 99.68]	98.66 [97.74, 99.46]	−0.63 [−1.23, −0.15]
Gemma-4-26B-A4B-it	DAPO-Math	79	99.37 [98.76, 99.77]	99.33 [98.54, 99.84]	−0.04 [−0.54, 0.34]
GLM-4.7-Flash	DAPO-Math	148	99.46 [99.13, 99.74]	98.98 [98.30, 99.55]	−0.48 [−0.99, −0.11]
Qwen3-30B-A3B	MATH-500	13	99.92 [99.77, 100.00]	99.85 [99.54, 100.00]	−0.08 [−0.23, 0.00]
Qwen3-235B-A22B	MATH-500	15	100.00 [100.00, 100.00]	100.00 [100.00, 100.00]	+0.00 [0.00, 0.00]
Gemma-4-26B-A4B-it	MATH-500	3	95.00 [85.00, 100.00]	96.00 [88.00, 100.00]	+1.00 [0.00, 3.00]
GLM-4.7-Flash	MATH-500	30	99.97 [99.90, 100.00]	100.00 [100.00, 100.00]	+0.03 [0.00, 0.10]
Qwen3-30B-A3B	GSM8K	22	99.45 [98.41, 100.00]	98.50 [95.68, 100.00]	−0.95 [−2.73, 0.00]
Qwen3.5-35B-A3B	GSM8K	6	98.67 [96.00, 100.00]	96.17 [88.50, 100.00]	−2.50 [−7.50, 0.00]
Qwen3-235B-A22B	GSM8K	25	99.16 [98.04, 99.96]	98.84 [97.40, 99.96]	−0.32 [−0.88, 0.00]
Gemma-4-26B-A4B-it	GSM8K	15	98.87 [97.20, 100.00]	97.27 [92.20, 100.00]	−1.60 [−5.80, 0.67]
GLM-4.7-Flash	GSM8K	38	99.87 [99.61, 100.00]	99.82 [99.45, 100.00]	−0.05 [−0.16, 0.00]
Qwen3-30B-A3B	HotpotQA	62	97.76 [96.63, 98.79]	97.18 [95.40, 98.68]	−0.58 [−1.60, 0.37]
Qwen3.5-35B-A3B	HotpotQA	48	98.46 [97.27, 99.40]	97.94 [96.33, 99.15]	−0.52 [−1.21, 0.10]
Qwen3-235B-A22B	HotpotQA	123	98.85 [98.22, 99.41]	98.11 [97.05, 99.11]	−0.75 [−1.31, −0.24]
Gemma-4-26B-A4B-it	HotpotQA	55	98.76 [97.85, 99.45]	96.73 [93.93, 98.69]	−2.04 [−3.82, −0.42]
GLM-4.7-Flash	HotpotQA	110	98.53 [97.76, 99.23]	98.06 [96.91, 98.96]	−0.46 [−1.14, 0.04]
Qwen3-30B-A3B	HumanEval	21	99.52 [98.57, 100.00]	99.00 [97.00, 100.00]	−0.52 [−1.57, 0.00]
Qwen3.5-35B-A3B	HumanEval	10	100.00 [100.00, 100.00]	100.00 [100.00, 100.00]	+0.00 [0.00, 0.00]
Qwen3-235B-A22B	HumanEval	16	98.31 [96.12, 100.00]	98.25 [96.25, 100.00]	−0.06 [−1.50, 1.50]
Gemma-4-26B-A4B-it	HumanEval	2	100.00 [100.00, 100.00]	100.00 [100.00, 100.00]	+0.00 [0.00, 0.00]
GLM-4.7-Flash	HumanEval	45	100.00 [100.00, 100.00]	100.00 [100.00, 100.00]	+0.00 [0.00, 0.00]

Table 22: **Complete Rare-retain@2 under synthetic scores.** S denotes Score-only and S+R denotes Score+routing with a shortlist of six. Metrics are percentages, and differences are percentage points. Brackets give 95% question-bootstrap intervals, paired for differences. †Fewer than five rare-correct questions, reported descriptively. Dashes indicate no rare-correct support.

Model	Dataset	Q_{rare}	S [95% CI]	S+R [95% CI]	Difference [95% CI]
Qwen3-30B-A3B	DAPO-Math	7	78.86 [75.57, 82.00]	73.43 [69.86, 77.14]	−5.43 [−10.29, −0.71]
Qwen3.5-35B-A3B	DAPO-Math	6	81.50 [76.17, 86.67]	75.17 [67.33, 81.67]	−6.33 [−11.33, −1.67]
Qwen3-235B-A22B	DAPO-Math	10	77.10 [74.80, 79.50]	66.40 [62.00, 71.50]	−10.70 [−14.90, −6.30]
Gemma-4-26B-A4B-it	DAPO-Math	3†	72.67 [71.00, 75.00]	82.00 [64.00, 94.00]	+9.33 [−8.00, 19.00]
GLM-4.7-Flash	DAPO-Math	8	77.50 [72.87, 82.25]	72.88 [69.12, 77.12]	−4.62 [−9.38, 0.37]
Qwen3-30B-A3B	MATH-500	0	–	–	–
Qwen3-235B-A22B	MATH-500	0	–	–	–
Gemma-4-26B-A4B-it	MATH-500	1†	68.00 [68.00, 68.00]	58.00 [58.00, 58.00]	−10.00 [−10.00, −10.00]
GLM-4.7-Flash	MATH-500	0	–	–	–
Qwen3-30B-A3B	GSM8K	1†	75.00 [75.00, 75.00]	58.00 [58.00, 58.00]	−17.00 [−17.00, −17.00]
Qwen3.5-35B-A3B	GSM8K	1†	80.00 [80.00, 80.00]	69.00 [69.00, 69.00]	−11.00 [−11.00, −11.00]
Qwen3-235B-A22B	GSM8K	2†	78.00 [73.00, 83.00]	74.00 [69.00, 79.00]	−4.00 [−14.00, 6.00]
Gemma-4-26B-A4B-it	GSM8K	2†	78.00 [76.00, 80.00]	63.00 [56.00, 70.00]	−15.00 [−20.00, −10.00]
GLM-4.7-Flash	GSM8K	1†	86.00 [86.00, 86.00]	70.00 [70.00, 70.00]	−16.00 [−16.00, −16.00]
Qwen3-30B-A3B	HotpotQA	13	75.77 [73.69, 77.69]	70.08 [66.62, 74.00]	−5.69 [−9.54, −1.54]
Qwen3.5-35B-A3B	HotpotQA	6	73.50 [69.83, 77.83]	69.17 [64.00, 75.33]	−4.33 [−12.02, 4.18]
Qwen3-235B-A22B	HotpotQA	12	75.42 [73.25, 77.58]	65.83 [62.42, 70.00]	−9.58 [−13.67, −5.67]
Gemma-4-26B-A4B-it	HotpotQA	7	77.57 [76.29, 78.86]	67.86 [61.29, 75.15]	−9.71 [−16.43, −2.43]
GLM-4.7-Flash	HotpotQA	14	75.86 [73.29, 78.57]	71.79 [65.50, 77.72]	−4.07 [−8.86, 0.71]
Qwen3-30B-A3B	HumanEval	1†	76.00 [76.00, 76.00]	55.00 [55.00, 55.00]	−21.00 [−21.00, −21.00]
Qwen3.5-35B-A3B	HumanEval	0	–	–	–
Qwen3-235B-A22B	HumanEval	3†	79.67 [68.00, 88.00]	78.33 [77.00, 80.00]	−1.33 [−8.00, 9.00]
Gemma-4-26B-A4B-it	HumanEval	0	–	–	–
GLM-4.7-Flash	HumanEval	0	–	–	–

E Representation and Parameter Diagnostics

We provide complementary interpretive observations of representation readout, local expert weighting, and router–expert parameters. The length-only baseline of this section is denoted B_{len}, distinguishing it from the candidate scorer S_B in Appendix D.

E.1 Four-way Representation Comparison

Representations. We replay the saved prompt and generation token IDs to extract router inputs and logits. For each selected layer and phase, we average over the same token span and concatenate the resulting vectors. The random projection uses orthonormal rows obtained from a Gaussian matrix by QR decomposition and has the same output dimension as the router logits. One fixed projection is shared across the compared trajectories and layers. Occupancy is selected from the archived phase tensors using the original rollout and layer indices. All four representations average within phases and therefore omit within-phase token order. At each selected layer, the router computes $z_l = W_l h_l$, while R projects the same input to the logit dimension. Hidden states and logits come from forward replay, while occupancy uses the archived routing records.

Readout and uncertainty. The baseline feature is $\log(1 + T)$, where T is the recorded `token_count`. We subtract the within-question mean from each feature without using correctness labels. Each question has total training weight one, distributed equally over its candidates. Feature scaling and regularization tuning use training folds, and all conditions share the same outer test questions. The length-only predictions are reused across the four comparisons. We compute AUC from candidate scores separately for each held-out mixed question and average equally across questions, and we denote this metric by M. This differs from the pair-order evaluation in §4, so numerical comparisons are restricted to this experiment. Confidence intervals resample these question-level values, using paired differences for comparisons. The intervals describe variation across evaluated questions and do not incorporate variation from new candidate sampling or new random projections.

Results. All four augmented predictors have lower AUC point estimates than the length-only baseline in this 50-question comparison (Table 23). Among the augmented predictors, the random projection has the highest AUC and router logits have the lowest, with hidden states and occupancy close to one another. The paired intervals in Table 24 describe the variability of these differences.

E.2 Gate Weights and Weighted Output Norms

For token t at layer l, let the selected experts have gate weights $\mathbf{w}$ and output norms $\mathbf{n}$. We compare

$$P_k^{\text{gate}} = \frac{w_k}{\sum_j w_j}, \qquad P_k^{\text{mag}} = \frac{w_k n_k}{\sum_j w_j n_j}. \tag{6}$$

The second distribution measures normalized magnitudes of the weighted expert outputs. It omits output directions, cancellation between experts, and downstream effects on correctness. For positive output norms,

$$D_{\text{KL}}(P^{\text{gate}} \| P^{\text{mag}}) = \log \frac{\sum_j w_j n_j}{\sum_j w_j} - \sum_k P_k^{\text{gate}} \log n_k. \tag{7}$$

The distributions coincide when output norms are equal over the experts with positive gate weights.

We analyze 50 rollouts per model–task setting in Table 25. Top-1 agreement records whether the largest gate weight and weighted output norm belong to the same expert. Top-3 agreement records whether the largest-gate expert is among the three largest weighted output norms. The credit ratio compares the gate probability with the magnitude probability at the largest-gate expert. Across the six settings, the cosine similarity between the two weightings ranges from 0.926 to 0.961, and the highest-gate expert falls among the three largest weighted output norms in more than 92% of positions. Gate weights and weighted output magnitudes therefore show substantial local agreement, providing a complementary description of selected experts.

Table 23: **Representation comparisons on 50 mixed questions from Qwen3-30B-A3B on DAPO-Math.** AUC measures within-question candidate ranking, with chance at 50%. AUC values are percentages and differences are percentage points. Intervals use paired question bootstrap for differences. The four augmented predictors share the same length-only baseline B_{len}.

Condition	AUC	95% CI	Δ vs. B_{len}	95% CI
B_{len}	62.19	$[54.51, 69.20]$	–	–
$B_{\text{len}} + h$	57.81	$[50.71, 64.91]$	-4.38	$[-12.50, 4.15]$
$B_{\text{len}} + Rh$	59.18	$[52.90, 65.35]$	-3.01	$[-10.68, 5.24]$
$B_{\text{len}} + z$	55.13	$[48.42, 61.61]$	-7.06	$[-16.36, 2.29]$
$B_{\text{len}} + \text{occupancy}$	57.93	$[51.65, 64.38]$	-4.26	$[-11.94, 3.83]$
Paired representation comparisons			Difference	95% CI
$M(B_{\text{len}} + Rh) - M(B_{\text{len}} + z)$			$+4.05$	$[-3.74, 11.10]$
$M(B_{\text{len}} + z) - M(B_{\text{len}} + \text{occupancy})$			-2.80	$[-9.95, 4.04]$

Table 24: **All paired representation differences.** Each predictor includes the shared length-only feature. Differences are in percentage points with 95% question-bootstrap intervals on the same 50 questions as Table 23.

Comparison	Difference	95% CI
$M(B_{\text{len}} + Rh) - M(B_{\text{len}} + z)$	$+4.05$	$[-3.74, 11.10]$
$M(B_{\text{len}} + z) - M(B_{\text{len}} + \text{occupancy})$	-2.80	$[-9.95, 4.04]$
$M(B_{\text{len}} + h) - M(B_{\text{len}} + \text{occupancy})$	-0.11	$[-7.52, 7.19]$
$M(B_{\text{len}} + h) - M(B_{\text{len}} + Rh)$	-1.37	$[-7.66, 5.08]$
$M(B_{\text{len}} + h) - M(B_{\text{len}} + z)$	$+2.68$	$[-6.29, 10.69]$
$M(B_{\text{len}} + \text{occupancy}) - M(B_{\text{len}} + Rh)$	-1.25	$[-7.99, 4.87]$

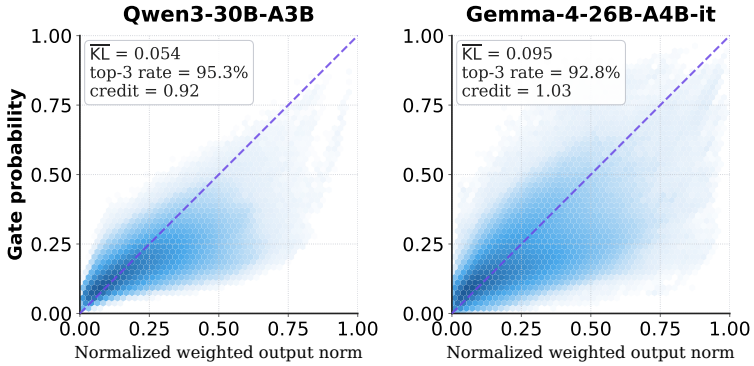


Figure 8: **Gate weights and weighted output norms.** The panels show pooled joint densities for Qwen3-30B-A3B on the left and Gemma-4-26B-A4B-it on the right. The horizontal axis gives normalized weighted output norms P^{mag}, the vertical axis gives gate probabilities P^{gate}, and the dashed line marks equality.

Table 25: **Gate and magnitude agreement by task.** Each row uses 50 rollouts. KL compares P^{gate} with P^{mag}, and top-1 and top-3 agreements are percentages.

Model	Dataset	KL	Cosine	Top-1	Top-3	Credit ratio
Qwen3-30B-A3B	MATH-500	0.062	0.953	63.3	94.6	0.922
	HotpotQA	0.049	0.961	65.6	95.8	0.921
	HumanEval	0.052	0.960	65.7	95.4	0.926
Gemma-4-26B-A4B-it	MATH-500	0.101	0.926	56.6	92.8	1.049
	HotpotQA	0.086	0.935	60.6	92.8	1.004
	HumanEval	0.098	0.928	58.9	92.9	1.034

E.3 Router–Expert Parameter Coupling

For layer l, let $W_l[i,:]$ be router row i and $G_{l,j} \in \mathbb{R}^{d_{\text{ff}} \times d}$ the gate-projection weights of expert j. We define

$$M_{ij} = \|W_l[i,:] G_{l,j}^{\top}\|_2, \qquad r_l = \frac{1}{E} \sum_{i=1}^{E} \frac{M_{ii}}{\sum_j M_{ij}}. \tag{8}$$

The statistic measures the share of the chosen interaction magnitude assigned to matching router–expert identities. The reference $1/E$ represents equal shares and is also the expected ratio under exchangeable expert identities. The comparison uses this equal-share reference to describe concentration in the parameter metric.

Figure 9 compares the five models and two additional Qwen3-30B checkpoints. The reported model-level ratios range from $0.85/E$ to $1.93/E$. The concentration under this metric varies across models and checkpoints.

E.4 Router Changes Across Checkpoints

We compare Base, instruction-tuned, and post-trained Qwen3-30B-A3B checkpoints using router-weight cosine and top-K membership changes. For the decision comparison, both router matrices act on the same reference hidden states. If $S_l^a(h_t)$ and $S_l^b(h_t)$ denote their selected expert sets, the reported flip rate is

$$\text{Flip}_l = \frac{1}{T} \sum_{t=1}^{T} \left(1 - \frac{|S_l^a(h_t) \cap S_l^b(h_t)|}{K} \right). \tag{9}$$

This measures the mean fraction of replaced expert memberships rather than the fraction of tokens with any change. Figure 10 shows high router-weight cosine with nonzero membership changes.

For $z_l = W_l h_l$, the change between two checkpoints satisfies

$$\Delta z_l = W_l \Delta h_l + \Delta W_l h_l + \Delta W_l \Delta h_l. \tag{10}$$

The router-change term obeys

$$\|\Delta W_l h_l\|_2 \le \|\Delta W_l\|_F \|h_l\|_2. \tag{11}$$

The shared-hidden-state comparison isolates changes in routing decisions induced by the router matrices. High weight cosine can coexist with membership changes because top-K selection depends on the ordering of nearby logits.

Together, these diagnostics describe representation readout, local expert weighting, and parameter variation.

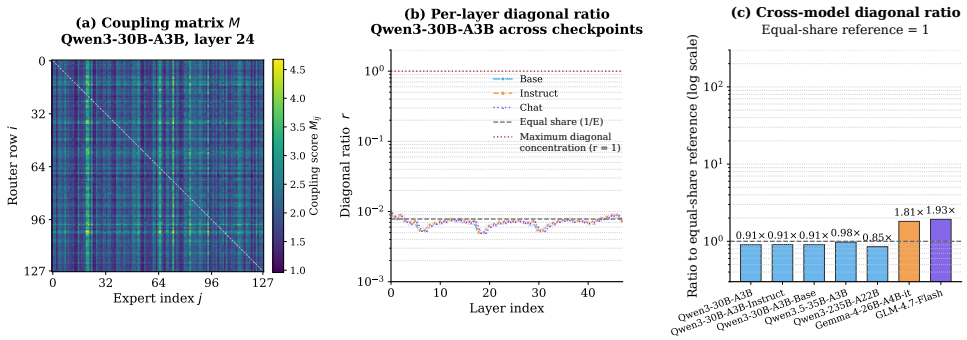


Figure 9: **Router–expert parameter coupling.** The panels show a representative Qwen3-30B-A3B coupling matrix, layer-wise diagonal ratios across its checkpoints, and a cross-model comparison. In panel (b), $1/E$ marks equal shares and $r = 1$ marks maximum diagonal concentration. Panel (c) reports ratios to the equal-share reference.

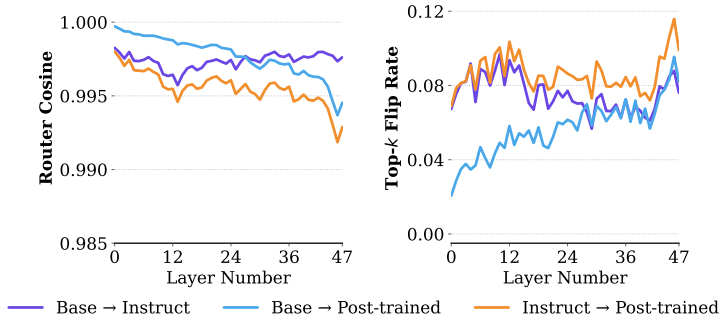


Figure 10: **Router parameters and decisions across checkpoints.** The panels show layer-wise router-weight cosine and mean top-K membership change for three Qwen3-30B-A3B checkpoint pairs. The decision comparison uses shared reference hidden states.

F Routing Profiles

F.1 Cross-task Profile Comparisons

Routing profiles describe where allocation changes across generation progress and how those changes vary with task and training. For each rollout and layer, we compare the expert-frequency distributions of the early, middle, and late progress bins defined in §2.2. We use the Jensen–Shannon distance (Menéndez et al., 1997),

$$d_{\mathrm{JS}}(p, q) = \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(p \| m) + \tfrac{1}{2} D_{\mathrm{KL}}(q \| m)}, \qquad m = \tfrac{1}{2}(p + q). \tag{12}$$

At each layer, we average the distances for early–middle, middle–late, and early–late within a rollout, then average across rollouts. The resulting profile measures how much expert allocation differs across generation progress, and we compare it across MATH-500, HotpotQA, and HumanEval within each of the five models.

Figure 11 shows layer-dependent profiles whose shapes are broadly similar across tasks within a model, with cross-task Pearson correlations from 0.62 to 0.96 (Table 26). Similar shapes coexist with different amplitudes, so the profile separates where allocation changes from how much it changes. For example, HotpotQA lies above MATH-500 for Qwen3-30B-A3B, while their ordering reverses for GLM-4.7-Flash. A phase-pair aggregation analysis finds correlations of 0.988–1.000 between adjacent-pair and all-pair profiles across the 24 available settings, indicating that the broad layer-wise pattern does not depend on one progress-bin pair (Table 27).

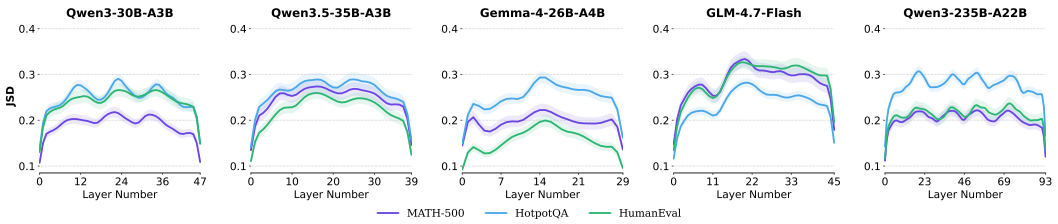


Figure 11: **Expert allocation changes across generation progress, with task-dependent magnitude.** Each curve is the mean JS distance across the three progress-bin pairs for one model and task, so higher values mark layers whose allocation differs more between progress bins. Curves within a model share a broad layer-wise shape, while their levels differ across tasks.

Table 26: **Cross-task similarity of layer-wise routing profiles.** Pearson correlations compare MATH-500, HotpotQA, and HumanEval within each model. These coefficients describe profile shape and do not measure agreement in absolute distance.

Model	Math/QA	Math/Code	QA/Code
Qwen3-30B-A3B	0.84	0.78	0.87
Qwen3.5-35B-A3B	0.96	0.96	0.93
Qwen3-235B-A22B	0.92	0.89	0.86
Gemma-4-26B-A4B-it	0.62	0.63	0.85
GLM-4.7-Flash	0.96	0.92	0.93

F.2 Phase-pair Aggregation

For rollout r and layer l, the distributions $p_{l,1}^{(r)}$, $p_{l,2}^{(r)}$, and $p_{l,3}^{(r)}$ correspond to the early, middle, and late phases in Section 2.2. The layer profile averages

$$J_l^{(r)} = \frac{1}{3}\left[d_{\mathrm{JS}}(p_{l,1}^{(r)}, p_{l,2}^{(r)}) + d_{\mathrm{JS}}(p_{l,2}^{(r)}, p_{l,3}^{(r)}) + d_{\mathrm{JS}}(p_{l,1}^{(r)}, p_{l,3}^{(r)})\right] \tag{13}$$

over available rollouts. The adjacent-pair alternative averages only the first two terms. Distances use natural logarithms, so their maximum is $\sqrt{\log 2}$. Both summaries average pairwise distances after taking the square root in Eq. 12.

Table 27 reports the phase-pair analysis on 24 available model–task settings in the archived routing data. For example, Qwen3-30B-A3B on DAPO-Math has an adjacent-pair mean of 0.1409 and an all-pair mean of 0.1514, while the two layer profiles correlate at 0.998. Across settings, the correlations range from 0.988 to 1.000. Including early–late generally raises the average distance while preserving a similar broad layer-wise pattern. These correlations compare aggregation rules within a setting and differ from the cross-task correlations in Table 26.

F.3 GSPO Training and Checkpoint Profiles

We train Qwen3-30B-A3B-Base with GSPO (Zheng et al., 2025) on DAPO-Math-17K (Yu et al., 2025). Training uses Adam (β_1=0.9, β_2=0.98) with a constant learning rate of 1×10^{-6} and weight decay 0.1. Each rollout batch draws 128 prompts with 8 responses per prompt (1,024 sequences), and we take 4 gradient updates per rollout batch. The maximum response length is 10,240 tokens. The policy objective uses clipped importance ratios with ϵ_{low}=0.2 and ϵ_{high}=0.28, with the KL coefficient set to 0.0. Rewards are computed by the DAPO rule-based verifier. Training was run on 16 H20 GPUs in a single VeRL setup, and we analyze checkpoints up to step 300. The routing-profile comparison evaluates MATH-500 generations from Base and checkpoints at steps 50, 150, 250, and 300. The analysis divides each generation into three equal-token bins and averages JS distance over the three bin pairs before averaging over rollouts. Each checkpoint generates its own continuations, so these profiles describe both the model state and its generated content. The comparison does not replay a fixed continuation through every checkpoint.

Table 27: **Phase-pair JS-distance comparisons.** Distances are averaged over rollouts and layers. Adjacent averages early–middle and middle–late, while All pairs additionally includes early–late. Profile r is the Pearson correlation between the two resulting layer profiles within the same setting.

Model	Dataset	Rollouts	Early–mid	Mid–late	Early–late	Adjacent	All pairs	Profile r
GLM-4.7-Flash	DAPO-Math	2231	0.1560	0.1468	0.1732	0.1514	0.1587	0.999
GLM-4.7-Flash	GSM8K	528	0.1825	0.2102	0.2184	0.1964	0.2037	0.998
GLM-4.7-Flash	HotpotQA	1718	0.2024	0.2019	0.2298	0.2021	0.2113	0.997
GLM-4.7-Flash	HumanEval	669	0.2322	0.2362	0.2614	0.2342	0.2432	1.000
GLM-4.7-Flash	MATH-500	456	0.2389	0.2565	0.2660	0.2477	0.2538	0.999
Gemma-4-26B-A4B-it	DAPO-Math	1073	0.1074	0.0923	0.1266	0.0998	0.1087	0.995
Gemma-4-26B-A4B-it	GSM8K	200	0.0917	0.0687	0.1008	0.0802	0.0870	0.996
Gemma-4-26B-A4B-it	HotpotQA	853	0.1616	0.1384	0.1845	0.1500	0.1615	0.997
Gemma-4-26B-A4B-it	HumanEval	32	0.1526	0.1584	0.1613	0.1555	0.1574	0.997
Gemma-4-26B-A4B-it	MATH-500	46	0.1108	0.0912	0.1280	0.1010	0.1100	0.991
Qwen3-235B-A22B	DAPO-Math	2508	0.1479	0.1352	0.1741	0.1415	0.1524	0.997
Qwen3-235B-A22B	GSM8K	391	0.1319	0.1181	0.1507	0.1250	0.1336	0.995
Qwen3-235B-A22B	HotpotQA	1873	0.2579	0.2332	0.2828	0.2456	0.2580	0.996
Qwen3-235B-A22B	HumanEval	231	0.1687	0.1583	0.2044	0.1635	0.1771	0.992
Qwen3-235B-A22B	MATH-500	217	0.1702	0.1581	0.1925	0.1642	0.1736	0.994
Qwen3-30B-A3B	DAPO-Math	3893	0.1486	0.1331	0.1726	0.1409	0.1514	0.998
Qwen3-30B-A3B	GSM8K	351	0.1467	0.1338	0.1718	0.1403	0.1508	0.997
Qwen3-30B-A3B	HotpotQA	975	0.2189	0.1875	0.2484	0.2032	0.2183	0.998
Qwen3-30B-A3B	HumanEval	349	0.2018	0.2092	0.2312	0.2055	0.2141	0.988
Qwen3-30B-A3B	MATH-500	199	0.1620	0.1504	0.1904	0.1562	0.1676	0.999
Qwen3.5-35B-A3B	DAPO-Math	317	0.1472	0.1231	0.1741	0.1351	0.1481	0.999
Qwen3.5-35B-A3B	GSM8K	66	0.1611	0.1273	0.1797	0.1442	0.1560	0.999
Qwen3.5-35B-A3B	HotpotQA	717	0.1913	0.1503	0.2181	0.1708	0.1866	0.999
Qwen3.5-35B-A3B	HumanEval	151	0.2066	0.1801	0.2320	0.1934	0.2063	0.998

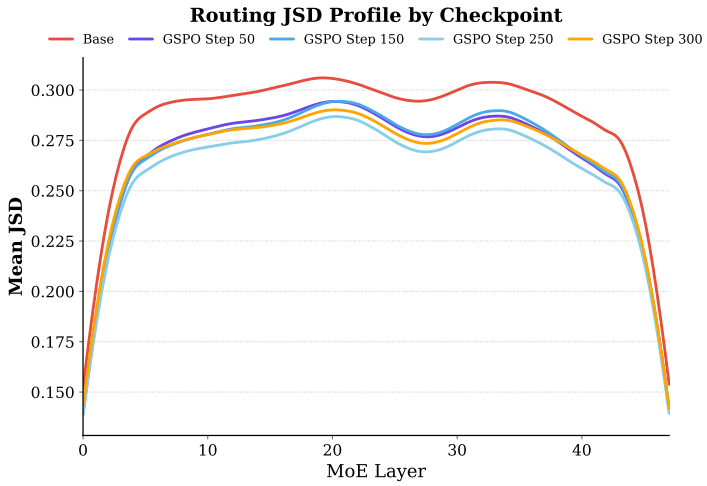


Figure 12: **Expert allocation structure persists while its level decreases during GSPO training.** JS-distance profiles compare equal-token progress bins on MATH-500. The displayed training checkpoints retain a similar layer-wise shape as Base while lying below it, indicating smaller between-bin allocation differences in the observed generations.

The displayed GSPO checkpoints retain a similar layer-wise shape at a lower level than Base (Fig. 12), indicating smaller allocation differences between progress bins in the observed generations. The profile therefore makes a change in allocation visible even when the broad layer-wise pattern persists. A separate checkpoint comparison shows high router-weight cosine alongside nonzero changes in selected expert membership (Figure 10).