SCREEN-THEN-SELECT: LEARNING WHEN TEMPERATURE SELECTION HELPS LLM REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sampling temperature controls the diversity of language-model reasoning, yet aggregate accuracy can conceal substantial differences in how inputs respond to temperature changes. We characterize this heterogeneity through multi-model MATH sweeps and high-sampling continuation experiments, connecting modest aggregate changes in a conventional temperature range to large local opportunities. Motivated by these findings, we propose Screen-Then-Select, an answer-assisted framework that separates screening for responsive candidates from selecting their temperatures. A small pool of reference-checked responses identifies mixed-success questions; a lightweight prompt-only selector then chooses a temperature for each fresh answer without updating the language model. In a MATH prefix study, correct-answer counts enrich inputs with large temperature responses. Trained on MATH and frozen for transfer, the selector improves individual-answer accuracy on screened GSM8K and OlympiadBench subsets relative to fixed temperatures selected for that metric on separate development questions, including a 7.1-percentage-point gain on GSM8K. Matched rejected controls show smaller observed gains, while policy decomposition identifies a shared temperature change as the source of the GSM8K improvement. Together, these results provide an empirical foundation and a lightweight framework for targeting temperature selection where its observed benefits are concentrated.

1 INTRODUCTION

Sampling temperature changes the reasoning paths a language model explores without changing its weights. Chain-of-thought prompting exposes intermediate reasoning (Wei et al., 2022), and self-consistency aggregates sampled solutions (Wang et al., 2023). Temperature therefore offers a direct way to adjust individual-answer accuracy and the answer distribution available for voting. Making effective use of this control requires understanding where temperature changes help.

Aggregate evaluations can suggest that tuning temperature offers limited benefit (Renze, 2024). Our multi-model MATH experiments likewise show modest aggregate changes over temperatures from 0.1 to 1.0. At the input level, however, adaptive decoding seeks to exploit differences across questions and reasoning states (Zhu et al., 2024; Su et al., 2026). This motivates the central question of our work: which inputs offer useful temperature-selection opportunities, and how can those opportunities be targeted?

We connect aggregate temperature response to local variation through high-sampling experiments. In six curated AIME prefixes, 1,024 continuations per temperature reveal higher-temperature benefit, lower-temperature benefit, and persistently low recovery. A larger MATH prefix experiment links response heterogeneity to the number of correct source answers. The 3–5/8 group contains 49.1% large-response prefixes, compared with 22.0% outside that group. This enrichment gives an empirical basis for screening responsive candidates before learning which temperature to select.

We propose Screen-Then-Select to implement this separation. An answer-assisted screen retains questions with three to five correct answers among eight samples. A MATH-trained MLP then reads only the question prompt, chooses one temperature, and holds it throughout each fresh response. The screen uses reference answers; the selector operates on prompt features alone. With the language model and selector frozen, the method yields positive individual-answer accuracy differences on screened GSM8K and OlympiadBench questions relative to independently development-selected

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.

fixed temperatures. Matched rejected controls and a gain decomposition explain where these benefits are concentrated and how shared temperature changes contribute to them.

Our contributions are:

1. **An empirical characterization of temperature-response heterogeneity.** Sweeps across three Qwen3 models and five MATH difficulty levels, together with high-sampling prefix experiments, reveal how modest aggregate changes coexist with substantial local temperature effects (Section 3).
2. **Correct-count screening for temperature-responsive candidates.** We demonstrate enrichment of large observed responses in a MATH prefix population and use disjoint response pools to examine repeatable selection opportunities. These findings motivate the screen-then-select design (Section 3.3).
3. **A lightweight framework with cross-task gains and attribution.** We introduce a frozen, prompt-only selector and demonstrate single-answer accuracy gains on screened target subsets. Rejected controls, common-temperature comparisons, and complete temperature sweeps distinguish the sources of these gains (Sections 4–5).

2 Related work

Aggregate temperature effects and answer aggregation. Renze (2024) finds no statistically significant aggregate temperature effect between 0 and 1 in a multiple-choice evaluation. Our free-form analysis separates that range from higher-temperature degradation and examines local response heterogeneity; it is not a direct replication or refutation.

For multi-sample inference, Du et al. (2025) select temperature using an entropy turning point without labeled task-specific validation; Li et al. (2025) study evolving answer distributions and confidence-driven calibration. Our answer-assisted screen incurs labeling and sampling cost. Individual accuracy and voting accuracy remain distinct endpoints.

Adaptive rules and learned temperature prediction. Nucleus sampling changes token eligibility (Holtzman et al., 2020); EDT maps uncertainty to temperature (Zhang et al., 2024), and AdapT uses token-distribution shape (Zhu et al., 2024). Our policy instead acts once before generation, using correctness-based targets across fixed temperatures.

TempNet learns input-dependent temperatures with a distributionally robust objective (Qiu et al., 2024); AutoDeco predicts temperature and top-p per token (Wang et al., 2025b). Input-dependent prediction is therefore not new. Neither is evaluated under our exact protocol, so no superiority is claimed.

Learning adaptive decoding and selecting useful states. Su et al. (2026) train prompt-level and token-level decoding policies with verifiable rewards and a frozen language model; their token-level training also filters sparse or noisy prompts. Our narrower contribution is to test screening enrichment and distinguish shared temperature changes from prompt-specific assignment using rejected controls and fixed-action comparisons.

Difficulty prediction versus temperature advantage. Correctness can be predicted from representations (Kadavath et al., 2022), and compute allocated by difficulty (Snell et al., 2025). Early-stopping self-consistency (Li et al., 2024) and difficulty-adaptive self-consistency (Wang et al., 2025a) adapt the sample budget, rather than identify which temperature helps. We center within-question targets and test transfer to GSM8K (Cobbe et al., 2021) and OlympiadBench (He et al., 2024), separating responsiveness from policy utility.

Selective prediction and screening. Selective question answering controls error by abstaining on uncertain inputs (Kamath et al., 2020). Our correctness-assisted screen instead defines a subgroup for temperature intervention; it is neither label-free abstention nor evidence of end-to-end cost savings. Rejected controls and fixed-action comparisons test whether this subgroup supports useful temperature selection.

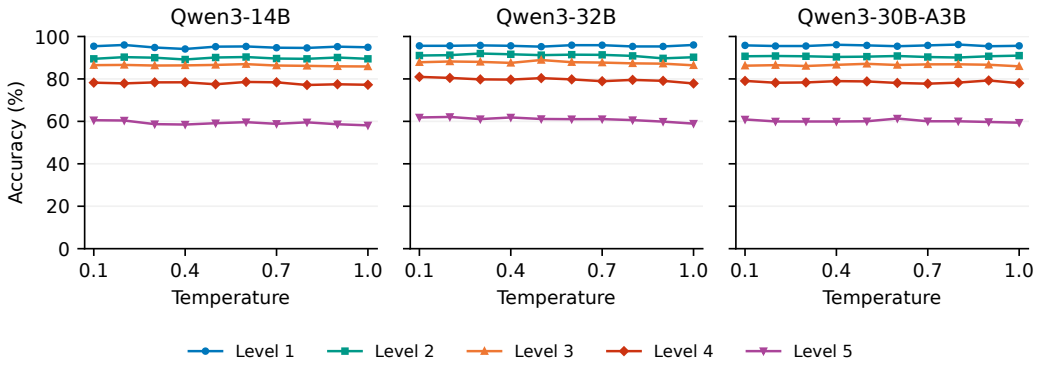


Figure 1: Aggregate temperature response on MATH for three Qwen3 models over $T = 0.1, 0.2, \ldots, 1.0$. Each panel shows five difficulty levels on the same accuracy scale. Across the 15 curves, the maximum-minus-minimum accuracy range is 0.80–3.14 percentage points. These are logged point estimates; aligned generation-level uncertainty is unavailable. Full-range curves, including high-temperature degradation, appear in Appendix A.

3 From Aggregate Temperature Response to Local Opportunity

3.1 Modest Aggregate Changes in the Conventional Temperature Range

We evaluate Qwen3-14B, Qwen3-32B, and Qwen3-30B-A3B (Yang et al., 2025) on MATH (Hendrycks et al., 2021) across five difficulty levels. Figure 1 focuses on $T \in [0.1, 1.0]$, the range central to reports of small aggregate temperature effects (Renze, 2024). The five levels contain 1,001, 2,241, 2,724, 2,904, and 3,629 evaluated rows, respectively. We measure individual-answer accuracy, not majority voting. Appendix A documents the question bank, recorded generation settings, and scoring protocol.

The 15 MATH model–level curves have maximum-minus-minimum accuracy ranges of only 0.80–3.14 percentage points in this interval. Thus aggregate evaluation can suggest limited benefit from adjusting temperature. This observation is range-specific: higher temperatures can substantially degrade accuracy, as the full sweeps in the appendix show. The question motivating our method is different: *can these modest aggregate changes conceal much larger gains on particular inputs?*

3.2 Three Local Response Patterns Hidden by Aggregation

An aggregate curve averages signed temperature effects: improvement on some inputs can be diluted by unaffected inputs or offset by deterioration elsewhere. To inspect this missing granularity, we evaluate six manually selected AIME2025 prefixes from Qwen3-8B chains whose final answers were wrong at $T = 0.1$. Each prefix has *1,024 continuations at each of 15 temperatures*, for 92,160 responses. Three prefixes were annotated as correct intermediate reasoning, while three already contained errors.

Figure 2 illustrates three patterns. The left example improves from 0.98% at $T = 0.1$ to 27.15% at $T = 1.1$. The middle example favors lower temperature: 44.63% at $T = 0.2$ versus 31.25% at $T = 0.9$. The right example remains below 5.38% throughout. For the same 0.1-to-0.9 change, the first gains 24.41 points while the second loses 8.98; averaging all six cases gives only +5.63 points. Local gains can therefore be diluted or offset in an average. The cases illustrate the mechanism, not its frequency in a benchmark. Recovery is also budget-dependent, especially for error-containing prefixes with frequent truncation. We next measure responsiveness and screening enrichment on a larger set of MATH prefixes.

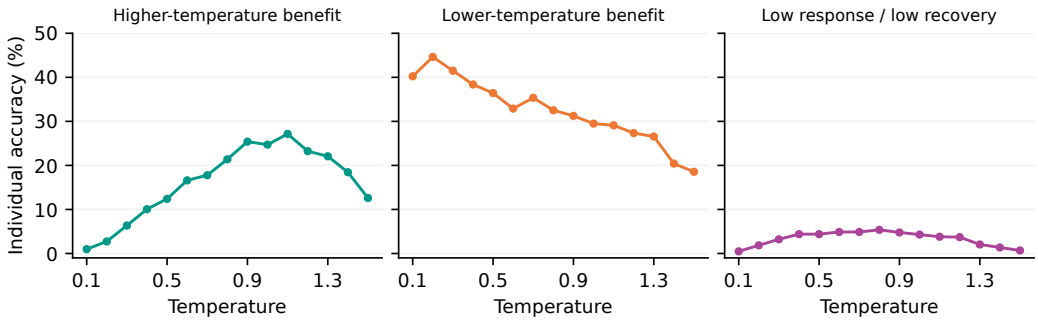


Figure 2: Three illustrative temperature-response patterns, using all 1,024 continuations at each temperature: higher-temperature benefit (left), lower-temperature benefit (middle), and persistently low recovery (right). The ordinate is the fraction of individually correct answers, not majority-vote accuracy. These curated examples do not estimate pattern prevalence. All six cases and split-pool checks are in Appendix B.

Table 1: Three checks of screening on the same MATH prefixes. **A:** observed large-response proportions versus a difficulty-preserving, temperature-independent null. The null entries are expected proportions, not p-values. **B–C:** choose actions on 64 answers per temperature and score on the other 64, averaging both directions over 2,000 splits. Gaps are screened minus outside in percentage points; 95% intervals bootstrap the 298 questions, not the splits. These are prefix diagnostics, not MLP results.

Measure	Screened (53)	Outside (295)	Gap [95% CI]
A. Are large responses enriched? (proportion, %)			
Observed range > 25 pp	49.06	22.03	+27.02 [10.94, 42.51]
Temperature-independent null	0.0424	0.0116	–
B. Do temperature differences repeat on separate answers? (pp)			
High minus low, $T = 0.1$–1.5	26.45	12.58	+13.87 [6.23, 21.35]
High minus low, $T = 0.1$–0.9	17.00	7.40	+9.60 [3.80, 15.48]
C. Can labeled answers identify an improvement over fixed? (pp)			
Selected best minus fixed 0.9	10.56	4.18	+6.38 [1.95, 11.33]

3.3 Correct-count screening enriches responsive inputs

For a fixed endpoint, let $u(x, T)$ be expected performance at input x (a question or a fixed prefix), and let $\mathcal{A}$ be the tested temperature grid. We distinguish three quantities:

$$\begin{aligned} \mathcal{R}(x) &= \max_{T\in\mathcal{A}} u(x,T) - \min_{T\in\mathcal{A}} u(x,T), && \text{response range,} \\ \mathcal{O}(x;T_{\text{ref}}) &= \max_{T\in\mathcal{A}} u(x,T) - u(x,T_{\text{ref}}), && \text{opportunity over fixed,} \\ g_\pi(x;T_{\text{ref}}) &= u(x,\pi(x)) - u(x,T_{\text{ref}}), && \text{selector gain.} \end{aligned} \tag{1}$$

Here u is individual accuracy. A large range can reflect harmful actions, whereas opportunity measures headroom above a fixed reference; selector gain can be negative. Policy evaluation uses answers separate from screening and fitting. SC@8 comparisons use voting accuracy instead.

We study 348 MATH prefixes from 298 questions, with 128 continuations at each of eight temperatures $\{0.1, 0.3, \ldots, 1.5\}$ and an eight-answer source probe at $T = 0.1$. The 3–5/8 rule retains 53 prefixes; the remaining 295 form the outside group. Despite a 17.48-point mean within-prefix range, the aggregate 0.1-to-0.9 change is just +0.61 points. Table 1 organizes the evidence into three questions.

A. Enrichment beyond sampling noise. Large observed ranges ($\widehat{\mathcal{R}} > 0.25$) are more common in the screened group. To check whether this only reflects the higher sampling variance of mixed-success inputs, we redistribute each prefix's pooled correct outcomes across temperatures while

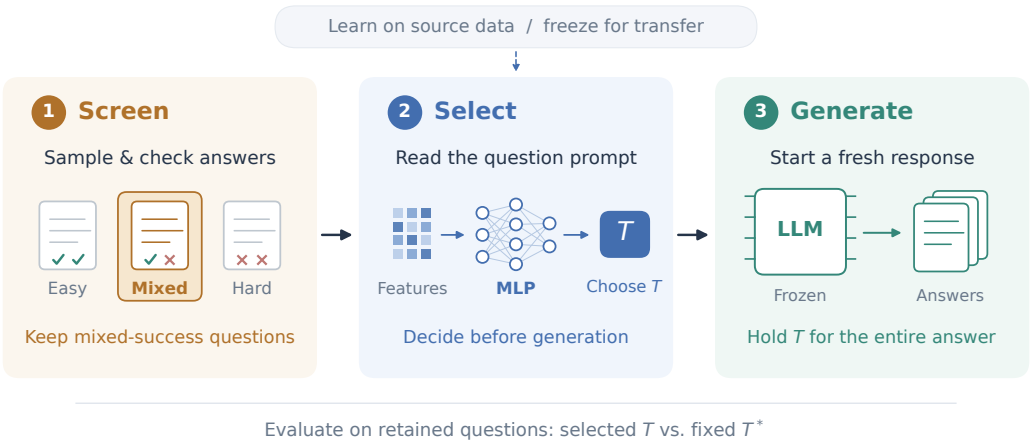


Figure 3: Screen-Then-Select overview. Screen: retain questions with 3–5 correct answers out of eight, checked against reference answers. Select: the MATH-trained, frozen MLP reads only the retained question prompt and chooses a temperature. Generate: use that temperature for each fresh complete answer and compare against a separately development-selected fixed temperature. Screening responses and correctness counts are not selector inputs. Outcome symbols are schematic, not literal sample counts.

preserving its total success count. Under this null, large ranges are rare in both groups. The observed enrichment is therefore not explained by binomial noise alone under the exchangeability assumption.

B. Repeatable temperature differences. Selecting high/low temperatures on one half of the answers and measuring their difference on the other retains the enrichment. It also persists when actions are restricted to $T \leq 0.9$, rather than depending only on harmful high temperatures.

C. Opportunity over a fixed reference. Choosing the best action on one half improves accuracy over fixed $T = 0.9$ on the other, with a larger gain in the screened group. This uses labeled answers to identify opportunities; it is not learned-selector performance. The 3–5/8 interval is a fixed mixed-success heuristic, not an established optimal threshold; Appendix C reports broader intervals and sensitivity checks. These fixed-prefix diagnostics motivate, but do not establish, the prompt-only policy gains tested separately in Section 5.

4 Screen-Then-Select

The preceding diagnostics distinguish response range from opportunities over a fixed reference. We implement screening and selection as a lightweight, prompt-only temperature policy (Figure 3).

4.1 Prompt-level decision problem

Let q denote the question together with the fixed prompt instructions. *Prompt-only selection* chooses a temperature before generating a fresh answer and applies it throughout. Let $R \in \{0, 1\}$ indicate final correctness, and define $Q(q, T) = \mathbb{E}[R \mid q, T]$. Our policy selects from the eight-action grid $\mathcal{A} = \{0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5\}$:

$$\widetilde{T}(q) = \arg\max_{T \in \mathcal{A}} \widehat{A}_\theta(q, T). \tag{2}$$

Here $\widehat{A}_\theta$ estimates centered success rates. The deployed policy π_θ applies the gate defined below to this candidate temperature and holds its chosen temperature throughout the answer.

4.2 ANSWER-ASSISTED SCREEN

For an independent screening pool of eight complete answers at $T = 0.1$, let $c(q)$ be the number of correct answers. We retain the intermediate-success set

$$\mathcal{S} = \{q : 3 \leq c(q) \leq 5\}. \tag{3}$$

Mixed observed success defines a candidate population on which to test temperature selection. Success at one temperature does not establish sensitivity to another: questions outside this set may also benefit from a temperature change. Screening uses reference answers to determine inclusion; its generated answers and correctness counts are excluded from the selector's inputs.

4.3 FROZEN MATH SELECTOR

The selector development cohort contains 350 MATH questions: 280 fit questions and 70 disjoint calibration questions. Its input is the mean of the final normalized layer-36 hidden states over the last 64 prompt tokens, concatenated with $\log(1 + n)$ and $n/8192$, where n is prompt length. Thus it has 4,098 input features and uses no PCA. It sees no generated answer, screen count, or gold label at decision time.

MATH-only feature development compared 19 cached *prefix* representations using best-vote-margin temperature classification. The selected layer-36 mean plus length is then used to train a *new prompt-only regressor*; the probe's classifier and cross-entropy results are not final-policy efficacy evidence.

Each fit/calibration question has 16 correctness labels per temperature. Their mean $\hat{p}(q, T)$ defines the target $a(q, T) = \hat{p}(q, T) - |\mathcal{A}|^{-1} \sum_{t \in \mathcal{A}} \hat{p}(q, t)$. A $4098 \to 64 \to 32 \to 8$ GELU MLP predicts all eight centered values, not a best-temperature class. We standardize features on fit questions only and minimize squared error after centering predictions across actions. AdamW training uses seed 42; calibration MSE selects the checkpoint, while calibration individual accuracy selects the learner and gate. The recorded checkpoint is epoch 5. Appendix D.2 gives optimization and label-provenance details. Centering removes the question-wide difficulty term, not the need to learn useful temperature differences. The policy uses $\widetilde{T}(q)$ only if $\widehat{A}_\theta(q, \widetilde{T}(q)) - \widehat{A}_\theta(q, T_0) > 0.05$; otherwise it uses the MATH-calibrated default $T_0 = 0.9$. This default remains frozen during transfer and is distinct from each target domain's fixed comparator T^*. The MLP, input normalization, action grid, and gate are frozen before all GSM8K and OlympiadBench evaluations.

Alternative decision timing. Prompt-only selection lets the selected temperature govern every response token. Generating a prefix before selecting instead changes both the available information and the partial answer. Appendix H separates prefix generation from subsequent selection in matched controls.

4.4 EVALUATION SETUP

Selector experiments use frozen Qwen3-8B and 280/70/45 MATH fit/calibration/ evaluation questions. The 3–5/8 screen retains 30/512 GSM8K and 56/512 OlympiadBench (He et al., 2024) questions, with same-size random rejected controls. For each target, a disjoint, unscreened 64-question development set with eight answers per temperature selects the individual-accuracy fixed comparator: $T^* = 1.1$ and 0.9, respectively. SC@8 selection instead gives $T^*_{\mathrm{SC}} = 0.9$ and 0.3. Individual accuracy is the selector's objective; Appendix D.5 records endpoint priorities across study stages.

Each target-domain policy uses eight fresh answers per question; MATH uses 32. Each answer is capped at 8,192 generated tokens. Actions are frozen before evaluation; equal actions share answers and different actions use separate seed namespaces. Screen, fit, development, and evaluation answers are separate; actual token usage need not match. We report equal-question individual accuracy, SC@8 (unique normalized-answer mode; ties wrong), and paired question-bootstrap intervals (Efron, 1979) (5,000 draws) from one rollout block. Appendices D.5, E.2, and D give the full protocol, grid, and development details; Appendix D.6 distinguishes the experimental populations.

Table 2: Matched rejected-control diagnostic. Controls are frozen random samples of the same size as screened subsets. Fixed baselines optimize development individual accuracy. Each Δ is selector minus fixed within that population; the enrichment gap is screened Δ minus rejected Δ, not a full-benchmark gain.

Dataset	Population	Δ Ind.	Δ SC@8	Δ Ind. 95% CI	Enrich. gap
GSM8K	screened (30)	+7.08	+6.67	[0.42, 13.75]	+5.42
GSM8K	rejected control (30)	+1.67	0.00	[-2.08, 5.42]	–
OlympiadBench	screened (56)	+1.56	+3.57	[-0.67, 4.46]	+2.90
OlympiadBench	rejected control (56)	-1.34	-1.79	[-3.79, 0.00]	–

Table 3: Prompt-only temperature selection by the frozen MATH MLP on screened target questions. Each policy uses eight completions per question. Each Δ is selector minus the target-domain development-selected fixed temperature; development selection maximizes individual-answer accuracy, not SC@8.

Dataset	Policy	Individual (%)	SC@8 (%)	Δ Ind.	Δ SC@8
GSM8K	fixed $T^* = 1.1$	58.33	76.67	–	–
GSM8K	MATH MLP	65.42	83.33	+7.08	+6.67
OlympiadBench	fixed $T^* = 0.9$	46.65	57.14	–	–
OlympiadBench	MATH MLP	48.21	60.71	+1.56	+3.57

5 Results

5.1 Screening concentrates individual-answer gains

Screened subsets show larger observed gains than same-size random rejected controls (Table 2). Using each task's individual-accuracy- selected fixed baseline, screened versus control gains are +7.08 versus +1.67 points on GSM8K and +1.56 versus -1.34 on OlympiadBench. The corresponding screen-minus-control gaps are +5.42 and +2.90 points.

Formally, for dataset d and population P, let $G_d(P) = \mathrm{Acc}_d(\text{selector}; P) - \mathrm{Acc}_d(\text{fixed}; P)$. The reported enrichment gap is $G_d(\text{screened}) - G_d(\text{rejected control})$; it is not a change relative to the complete benchmark.

Appendix E.2 gives absolute accuracies and control curves; Section 5.3 examines voting-specific baselines.

5.2 Transfer evaluation against development-selected fixed temperatures

The MATH-trained selector improves individual-answer accuracy on both screened target subsets without target-domain fitting (Table 3). On 30 GSM8K questions it gains 7.08 percentage points over the development-selected fixed baseline, with paired 95% interval $[+0.42, +13.75]$. On 56 OlympiadBench questions the gain is +1.56 points, with an interval including zero. Section 5.3 examines endpoint-matched voting baselines.

The MATH representation, MLP, and gate remain frozen during transfer. Target reference answers define the screen, not selector inputs or fitting.

On the 45-question MATH evaluation split, both the prompt-only MLP and fixed $T = 0.9$ obtain 54.17% individual accuracy. A post-hoc partition of each policy's 32 responses into four fixed eight-answer blocks gives 68.33% versus 66.67% SC@8 (+1.67 points). All three net block wins occur on one question; the 180 question–blocks are not independent questions. Appendix E.1 gives the source-domain results.

Source of the observed gains. The selector uses $T = 0.9$ for all 30 screened GSM8K questions, so its gain is a shared temperature change rather than heterogeneous question-level selection.

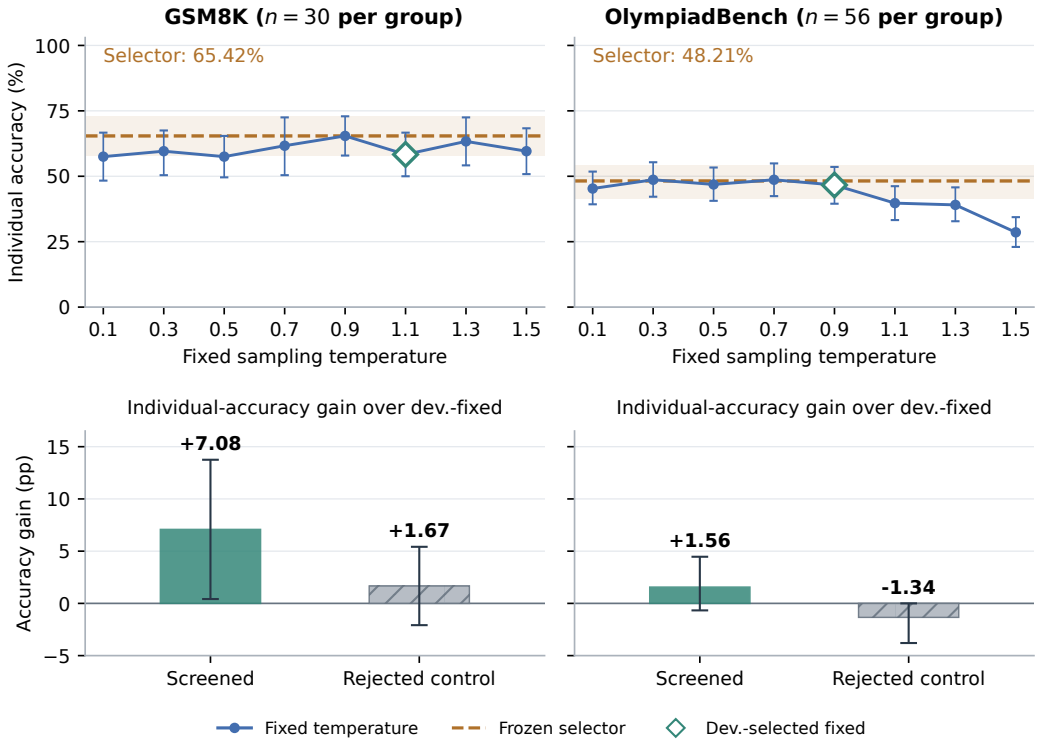


Figure 4: Frozen temperature selection and screening enrichment, with eight answers per question and temperature. **Top:** individual-answer accuracy across the complete fixed-temperature grid on screened questions, with the frozen selector dashed and the development-selected fixed baseline marked by a diamond. The GSM8K selector coincides with $T = 0.9$ because identical actions share responses. **Bottom:** selector-minus-fixed gains on screened questions and same-size random rejected controls, using the same development-selected baseline within each dataset. Error bars and upper-row shaded bands are pointwise 95% question-bootstrap intervals conditional on saved response pools; lower-row intervals are for paired policy differences, not the between-group gain gap. Both tasks' SC@8 curves appear in Appendix E.2.

On OlympiadBench it changes five actions from the default 0.9: three to 0.5 and two to 0.3. Appendix E.4 gives the complete accounting decomposition and rejected-control comparison.

5.3 Full-grid comparisons and baseline sensitivity

The frozen MATH selector matches GSM8K's observed fixed-grid maximum of 65.42% at $T = 0.9$ and reaches 48.21% on OlympiadBench, only 0.45 percentage points below its maximum of 48.66% at $T = 0.3$ and 0.7 (Figure 4, top). These maxima are post-hoc references, not the independently development-selected comparators used for the +7.08 and +1.56-point gains. No target-domain refitting or action changes are made during grid completion.

Screening concentrates the transfer gains. Figure 4 (bottom) connects this grid comparison to the larger gains on screened questions than on rejected controls (Table 2), showing where the benefits are concentrated.

Majority voting is a separate endpoint. The individual-answer gains do not extend to superiority over fixed baselines selected for SC@8: the differences are 0.00 points on GSM8K and -10.71 on OlympiadBench (60.71% versus 71.43%; paired 95% interval $[-25.00, +3.57]$ points). Figure 12 and Appendix E.2 retain both tasks' full SC@8 curves, endpoint-matched baselines, and rejected-control comparisons.

Screened-development fixed comparison. Applying the screen to the separate development sets retains only 2 GSM8K and 9 OlympiadBench questions. Selecting a fixed temperature among the seven actions not used for screening gives $T = 1.3$ and 0.9, respectively; selector gains become +2.08 and +1.56 points. This retrospective comparison preserves positive point estimates but shows baseline sensitivity, particularly with only two GSM8K development questions (Appendix E.3).

6 DISCUSSION

Screening makes response heterogeneity useful. Difficulty-preserving randomization and disjoint-half evaluation support response enrichment in MATH prefixes. Larger observed policy gains on screened target questions connect this motivation to practical selection: a shared temperature change on GSM8K and a few prompt-specific changes on OlympiadBench.

A lightweight route to decoding improvement. The frozen prompt-only selector requires neither target-domain fitting nor language-model weight updates. Separating screening from advantage prediction complements adaptive decoding (Su et al., 2026) with explicit population selection.

7 LIMITATIONS

Dependence on correctness feedback. The screen uses reference answers to assess pilot responses and is therefore answer-assisted rather than a label-free deployment gate. It cannot be directly applied to general open-ended tasks without reference answers.

Additional inference cost. Screening requires eight pilot responses per question. Deployment must account for this overhead and compare allocating the same inference budget to screening versus directly generating additional answers.

Task- and model-dependent screening. The fixed 3–5/8 rule may require adaptation: the same correct-answer count need not indicate the same temperature-selection opportunity across models of different capabilities or tasks of different difficulty. The threshold's transferability requires further study.

Restricted control space. The selector chooses one temperature from a discrete grid and holds it throughout the answer. This design does not respond to changing generation states or jointly optimize top-p, sampling budgets, or other decoding factors.

8 CONCLUSION

Temperature-response heterogeneity creates opportunities hidden by benchmark averages. Our multi-model sweeps and high-sampling experiments characterize this variation, while correct-count screening enriches large responses in the studied MATH prefix population. Screen-Then-Select improves individual-answer accuracy on screened transfer subsets relative to development-selected fixed temperatures. Distinguishing shared from question-specific temperature choices provides a foundation for screening-guided decoding optimization.

AI USE STATEMENT

Generative AI assistance was used in developing and discussing research hypotheses and evaluation methodology, implementing and inspecting analysis code, interpreting recorded results, identifying literature, and drafting and editing this manuscript. AI assistance was also used to organize references, prepare the LaTeX source, and generate figure-rendering code from saved numerical summaries. The evaluated language-model continuations are themselves model-generated research data. The reported experiments include fresh, frozen-policy prompt-only evaluations but no target-domain selector training. The authors have reviewed the manuscript and take responsibility for its content and the use of AI assistance.

Ethics statement

The reported task uses mathematical reasoning problems and generated answers, not a human-subject experiment. Correctness-based screening restricts the evaluation population and can make reported performance appear stronger than unrestricted performance if that restriction is omitted. We explicitly state the screening information, population counts, baseline-selection rule, and limitations of the current one-rollout evidence to reduce this risk. No use in safety-critical decision-making is established. Any eventual release of model outputs and dataset-derived material must respect the applicable model and dataset terms; the anonymous supplementary materials contain code, selector artifacts, derived outcomes, and a minimal normalized-answer export, but no raw benchmark problems, generated reasoning traces, or third-party language-model weights.

Reproducibility statement

The anonymous code supplement provides the frozen MATH selector and feature transform, screening and question-role records, per-question evaluation outcomes, generation and scoring source code, and a checksum manifest. Its CPU-only entry point, `reproduce_tables.py`, reconstructs the MATH individual-accuracy and four-block SC@8 results, both target-domain selector/fixed comparisons and rejected controls, and the complete target-domain fixed-temperature curves from the supplied outcomes. It does not rerun model generation, semantic grading, or confidence-interval estimation. The accompanying paper source contains numerical aggregates and plotting code for the descriptive analyses, including the selected W&B accuracy histories.

Sections 4–4.4 define the policy and endpoints; Appendix D records training, dataset construction, generation, scoring, and the boundaries between experimental populations. Appendix D.8 specifies the supplement's entry points and scope. Full regeneration requires the public benchmarks, matching Qwen3-8B weights, and the documented software environment. Raw reasoning traces and the complete training-label pool are not distributed; the supplement supports numerical replay and inspection of the frozen policy rather than a turnkey end-to-end rerun.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.neurips.cc/paper_files/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Weihua Du, Yiming Yang, and Sean Welleck. Optimizing temperature for language models with multi-sample inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14648–14668. PMLR, 2025. URL https://proceedings.mlr.press/v267/du25f.html.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset.

In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. doi: 10.2307/4615733. URL https://www.jstor.org/stable/4615733.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1904.09751.

Saurav Kadavath, Tom Conerly, Amanda Askell, et al. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221, 2022. URL https://arxiv.org/abs/2207.05221.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.503. URL https://aclanthology.org/2020.acl-main.503/.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/3fe2a777282299ecb4f9e7ebb531f0ab-Abstract-Conference.html.

Yiwei Li, Ji Zhang, Shaoxiong Feng, Peiwen Yuan, Xinglin Wang, Jiayi Shi, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. Revisiting self-consistency from dynamic distributional alignment perspective on answer aggregation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25208–25223. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1293. URL https://aclanthology.org/2025.findings-acl.1293/.

Belinda Phipson and Gordon K. Smyth. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9:Article 39, 2010. doi: 10.2202/1544-6115.1585. URL https://doi.org/10.2202/1544-6115.1585.

Zi-Hao Qiu, Siqi Guo, Mao Xu, Tuo Zhao, Lijun Zhang, and Tianbao Yang. To cool or not to cool? temperature network meets large foundation models via DRO. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 41604–41643. PMLR, 2024. URL https://proceedings.mlr.press/v235/qiu24c.html.

Matthew Renze. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7346–7356, 2024. doi: 10.18653/v1/2024.findings-emnlp.432. URL https://aclanthology.org/2024.findings-emnlp.432/.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Chloe H. Su, Zhe Ye, Samuel Tenka, Aidan Yang, Soonho Kong, and Udaya Ghai. Learning adaptive LLM decoding. arXiv preprint arXiv:2603.09065, 2026. URL https://arxiv.org/abs/2603.09065.

Xinglin Wang, Shaoxiong Feng, Yiwei Li, Peiwen Yuan, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. Make every penny count: Difficulty-adaptive self-consistency for cost-efficient reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 6919–6932. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.findings-naacl.383. URL https://aclanthology.org/2025.findings-naacl.383/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2203.11171. URL https://arxiv.org/abs/2203.11171.

Zhichao Wang, Dongyang Ma, Xinting Huang, Deng Cai, Tian Lan, Jiahao Xu, Haitao Mi, Xiaoying Tang, and Yan Wang. The end of manual decoding: Towards truly end-to-end language models. arXiv preprint arXiv:2510.26697, 2025b. URL https://arxiv.org/abs/2510.26697.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Shimao Zhang, Yu Bao, and Shujian Huang. EDT: Improving large language models' generation by entropy-based dynamic temperature sampling. arXiv preprint arXiv:2403.14541, 2024. URL https://arxiv.org/abs/2403.14541.

Yuqi Zhu, Jia Li, Ge Li, YunFei Zhao, Jia Li, Zhi Jin, and Hong Mei. Hot or cold? adaptive temperature sampling for code generation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. doi: 10.48550/arXiv.2309.02772. URL https://arxiv.org/abs/2309.02772.

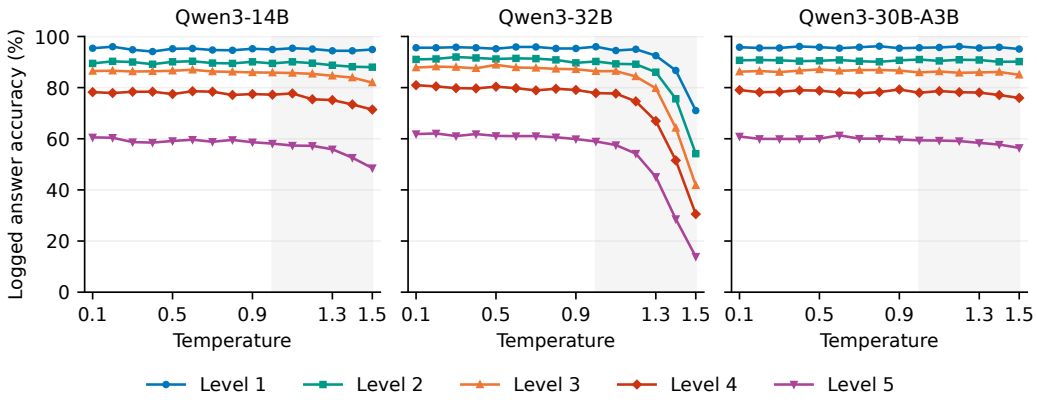


Figure 5: Complete MATH temperature curves for all three models and five difficulty levels. Shading distinguishes $T > 1$ from the main-text reporting range; it is not a fitted change point. All 225 measurements are shown.

Table 4: Accuracy at $T = 0.1 \rightarrow 1.5$ (change in points). MATH uses level 5 with 3,629 rows; AIME uses 90 questions and flexible answer extraction. Task-specific generation and grading differ, so the columns are not a controlled comparison of task difficulty.

Model	MATH level 5	AIME
Qwen3-14B	60.54 → 48.53 (-12.01)	4.44 → 3.33 (-1.11)
Qwen3-32B	61.84 → 13.89 (-47.95)	10.00 → 1.11 (-8.89)
Qwen3-30B-A3B	60.90 → 56.43 (-4.46)	4.44 → 3.33 (-1.11)

Appendix guide. Appendices A–C provide temperature-response and screening evidence. Appendices D–E give the final selector's protocol and complete results. Appendices F–H document supervision diagnostics, alternative selectors, and prefix-based controls.

A Multi-model experiments: protocols and full-range results

Full-range temperature response. Our complete MATH sweeps cover $T = 0.1, 0.2, \ldots, 1.5$, giving 225 model–level–temperature measurements (Figure 5). The small aggregate changes emphasized in the main text apply to 0.1–1.0, not the whole grid. On level 5, increasing temperature from 0.1 to 1.5 reduces accuracy by 12.01, 47.95, and 4.46 points for Qwen3-14B, Qwen3-32B, and Qwen3-30B-A3B. Table 4 reports endpoints for both tasks; Table 5 gives all MATH strata. High-temperature degradation differs across models and need not increase monotonically with difficulty. The separate MoE extension in Figure 7 shows that stability does not continue indefinitely beyond the main-text range.

Six-model AIME extension. The earlier local response dataset covers Qwen3-0.6B, 4B, 8B, 14B, 32B, and 30B-A3B on the same 90 AIME questions, each with one saved response at each of 15 temperatures from 0.1 to 1.5 (8,100 responses). Each model–temperature point uses 90 answers, giving 1,350 per model. One changed correctness outcome shifts accuracy by 1.11 percentage points. Single-generation sampling does not separate repeat-generation variability from temperature effects. This supplementary sweep is not the main-text multi-model evidence. Question IDs and exact question texts were aligned to the AI-MO AIME validation dataset. This dataset is separate from the AIME2025 WSP cases below. The local generation source uses a plain step-by-step prompt, a 2,048-output-token cap, unrestricted top-k, and top-$p = 1$. It does not use the later Qwen chat template. Complete generation-time logs and finish reasons are unavailable, so a low score cannot be attributed uniquely to model ability or truncation.

We regrade saved response steps using a gold-independent extractor followed by bounded symbolic equivalence. The flexible extractor tries a final box, a conclusion phrase, terminal mathematics,

Table 5: Complete MATH endpoint comparisons for all 15 model–level curves. Accuracy is in percent; differences are percentage points. Logged n counts rows, not independent generations per question.

Qwen3	Level	Logged n	$T = 0.1$	$T = 1.0$	$T = 1.5$	$\Delta_{1.5-0.1}$
14B	1	1,001	95.40	94.91	94.91	-0.50
14B	2	2,241	89.47	89.47	88.00	-1.47
14B	3	2,724	86.49	85.87	81.98	-4.52
14B	4	2,904	78.24	77.24	71.42	-6.82
14B	5	3,629	60.54	58.12	48.53	-12.01
32B	1	1,001	95.60	96.00	71.03	-24.58
32B	2	2,241	91.03	90.23	54.17	-36.86
32B	3	2,724	87.85	86.42	41.81	-46.04
32B	4	2,904	80.96	77.86	30.54	-50.41
32B	5	3,629	61.84	58.94	13.89	-47.95
30B-A3B	1	1,001	95.80	95.60	95.10	-0.70
30B-A3B	2	2,241	90.67	90.99	90.18	-0.49
30B-A3B	3	2,724	86.23	85.94	85.02	-1.21
30B-A3B	4	2,904	79.06	78.00	75.96	-3.10
30B-A3B	5	3,629	60.90	59.38	56.43	-4.46

and short final text; it does not search for the reference answer. Missing answers and verification failures count as incorrect. Because the original prompt does not require a boxed answer, Figure 6 reports both flexible extraction and the more restrictive last-closed-box rule. At $T = 0.1 \rightarrow 1.5$, box-only accuracy is 2.22%→1.11% for 14B, 10.00%→0.00% for 32B, and 0.00%→2.22% for 30B-A3B. Flexible-extraction endpoints are in Table 4. The extraction sensitivity, especially for the MoE model, rules out a strong model-ranking claim from these data. Question-bootstrap intervals condition on saved generations; they do not measure repeated-generation uncertainty. We preserve all six model curves, including flat and irregular ones, without treating their maxima as validated temperature choices.

Run selection and coverage. Our W&B experiment records contain 65 runs: 33 with an accuracy history, 30 with process metrics but no accuracy history, and two without usable history. Fifteen form the main cohort. Selection requires the full 0.1–1.5 grid, one generator per run, `subset_size=0`, `max_model_len=8192`, batch size 1, and the 34B process-evaluator setting. All five levels are available for Qwen3-14B, Qwen3-32B, and Qwen3-30B-A3B, with constant per-level logged counts across temperatures. This rule uses configuration and coverage, not a favorable outcome. Five separate Qwen3-30B-A3B runs cover 1.6–2.5. The remaining 13 accuracy-bearing runs have other configurations, incomplete coverage, or overlapping model–level conditions and are not pooled as independent repeats. In particular, earlier Qwen3-8B curves do not belong to the matched main cohort. The anonymous aggregate includes the complete 65-run inventory and the 225 main and 50 extension points.

Question-bank scope. The multi-model sweep characterizes temperature response using a MATH question bank spanning the dataset's original partitions. These original dataset partitions are distinct from the question-disjoint fit, calibration, and evaluation splits used to train and assess the selector (Appendix D). Local level files matching the logged counts contain questions from both original partitions. Levels 1–5 contain respectively 1,001, 2,241, 2,724, 2,904, and 3,629 rows, totaling 12,499. The level-5 file has 3,628 unique question IDs: one ID occurs twice. We preserve recorded aggregates rather than silently deduplicate a different response pool. These counts describe the available bank; the complete row-level identity of every logged run has not been independently reconstructed. Thus this analysis is neither a clean MATH test-split result nor an independent sample of 562,455 questions. The latter number is only the nominal sum of logged coverage over 225 points.

Metric definition and recorded generation settings. The recorded source revision extracts the last closed boxed answer and computes a MATH-style normalized-string equivalence flag against the reference. The logged `accuracy` averages these flags; it is distinct from ReasonEval validity

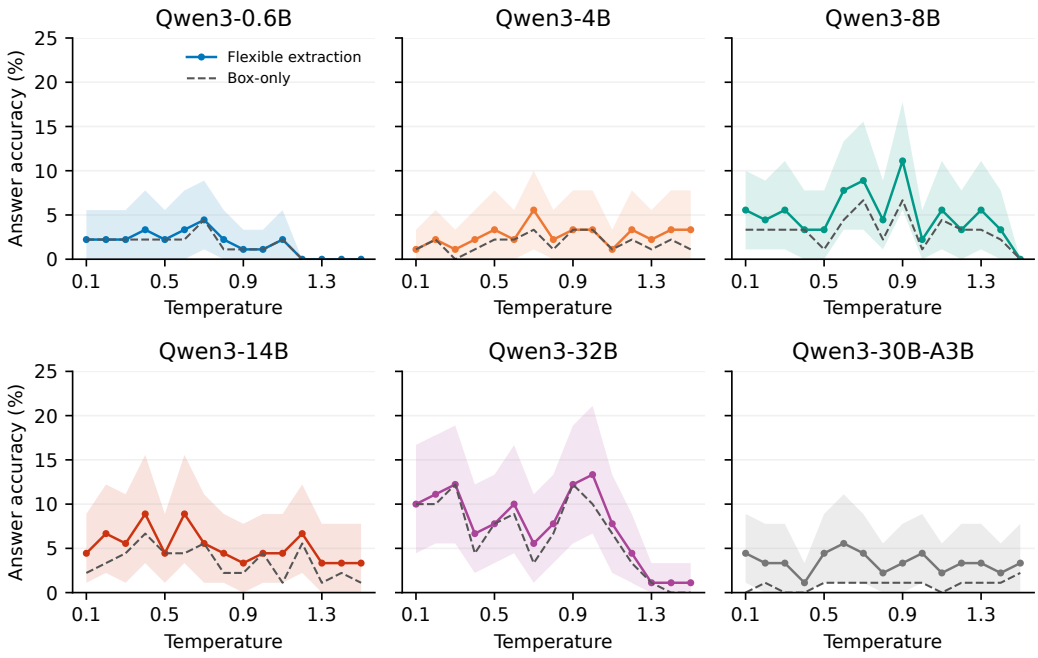


Figure 6: Six models on 90 aligned AIME questions under the older plain-prompt protocol. Colored curves use flexible extraction; dashed curves use boxes only. Shading is a pointwise 95% question-bootstrap interval for flexible extraction. All 90 temperature points are shown. These curves do not share the generation or grading protocol of the MATH dataset and are not selector replications.

and Math-Shepherd scores. The code's `num_samples` counts successful process evaluations, not repeated generations per question. The associated generator retains one completion per input and uses a chat prompt with thinking disabled, $p = 1$ and unrestricted top-k. This differs from the earlier local plain-prompt dataset and the later selector experiments.

Data source and reporting scope. The multi-model curves directly use the accuracy values recorded in the selected W&B runs. Because this analysis has not completed response-level matching to those run records, we report the recorded point estimates without constructing additional question-paired statistics or comparisons of actual token expenditure. No grid peak is treated as validation-selected, and no inference of per-question temperature advantage follows from these aggregate curves.

Relation to the original analysis. The earlier written analysis emphasized approximate stability in the conventional low-temperature range and uncertainty in process evaluation. The complete accuracy-bearing dataset refines that account: a low-range plateau can coexist with model- and difficulty-dependent degradation at higher temperatures. It does not justify the stronger statement that increasing temperature always hurts, or that harder questions always have a larger temperature effect. Figures 7 and 8 retain the extension and auxiliary metrics separately from the main comparison.

B WSP HIGH-SAMPLING CASE STUDY AND SPLIT-POOL REPLAY

Case selection and generation. WSP denotes a wrong-source-prefix case, not a benchmark difficulty level. All six source responses were final-answer incorrect at $T = 0.1$. Prefixes were cut at approximately one quarter of the source trajectory in 64-token segments, then manually annotated. WSP1–3 were judged correct up to the cut (modular enumeration, lattice counting, and a telescoping product). WSP4–6 contained an unjustified geometry assumption, misread segment distances, and an incorrect perimeter equation. The annotation labels are descriptive; the continuation results do

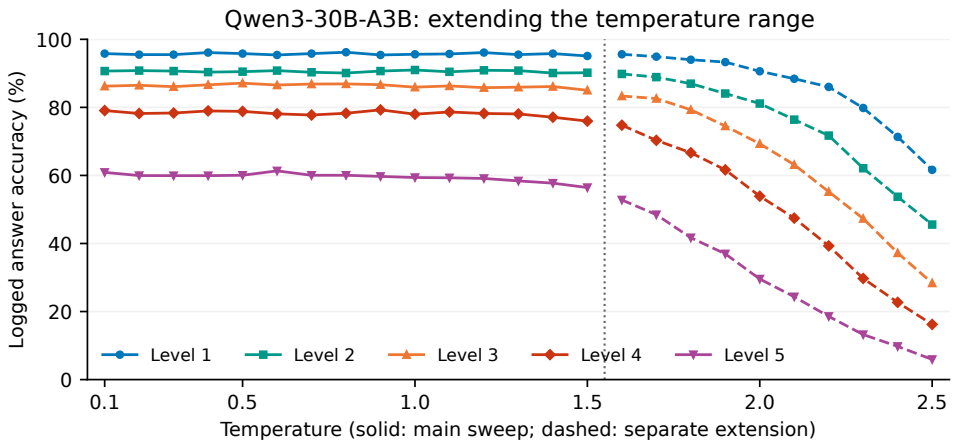


Figure 7: A separate high-temperature extension of Qwen3-30B-A3B across the same five named MATH strata. Solid and dashed segments are different recorded runs, separated between 1.5 and 1.6; they are not independent replications of one another. All extension points are shown. Relative stability on 0.1–1.5 does not imply temperature invariance outside that range.

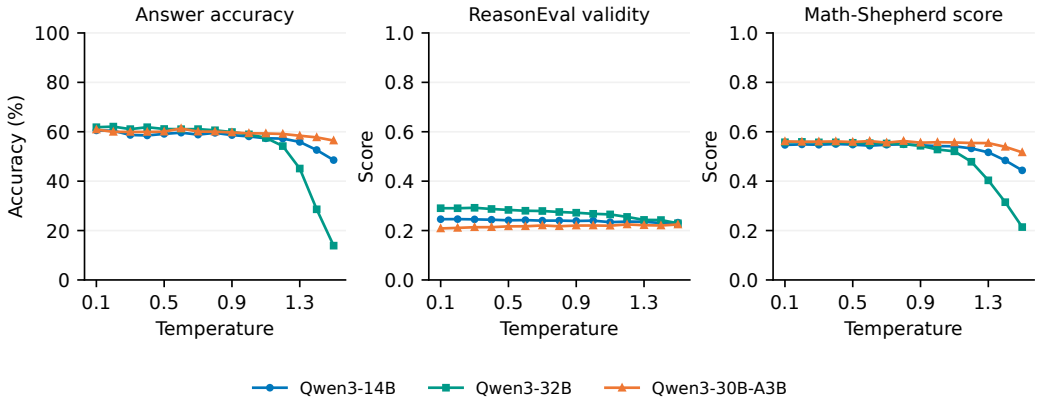


Figure 8: Answer accuracy and process scores on level 5 for the same three main-cohort models. The three panels have different metric definitions and scales. Process scores are auxiliary measurements, not correctness probabilities or substitutes for answer accuracy. Their trend differences motivate preserving the endpoint distinction.

not identify a causal effect of prefix correctness. Cases were curated, not sampled to estimate a population fraction.

For each fixed prefix, Qwen3-8B generates 1,024 continuations at each of 15 temperatures, totaling 15,360 per case and 92,160 overall. The complete 8192-budget runs are used; an interrupted longer-budget attempt is excluded. Generation uses the saved chat-rendered prompt and prefix, top-$p = 1$, unrestricted top-k, and at most 8,192 tokens including the prefix. A request produces 32 responses; 1,024 answers therefore do not denote 1,024 separate training runs or independent request seeds. Recorded correctness flags are retained, and their means exactly reproduce all 90 recorded case–temperature entries. A simple integer audit differs on 190 correctly scored algebraic answer forms, such as $21 + 49 = 70$ and $m + n = 106$; these are not converted to errors merely because a direct floating-point parser cannot read them.

Disjoint-pool replay and its limits. Within each case and temperature, the first 512 ordered responses select T_A by individual accuracy; the last 512 evaluate it. Whole 32-response requests remain within a pool. Ties select the lower temperature. Table 6 shows both fixed references for

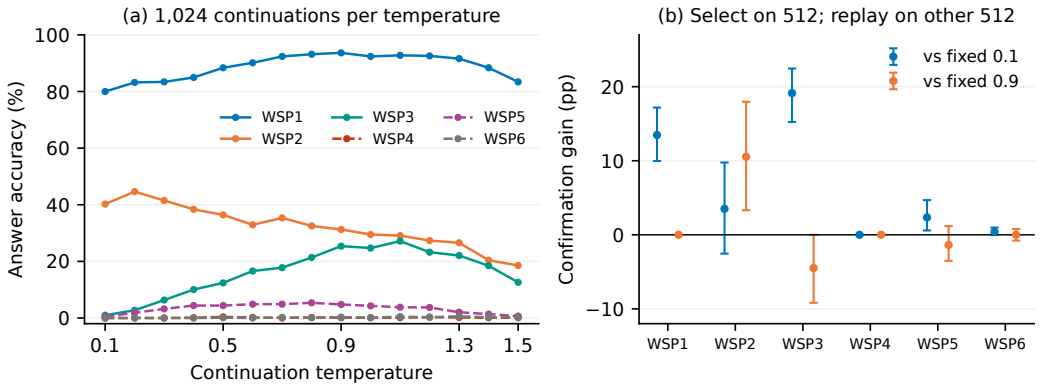


Figure 9: All six 1,024-response curves and retrospective 512/512 replay. Solid lines mark prefixes annotated as correct, dashed lines erroneous prefixes. Replay intervals are pointwise 95% request-block bootstrap intervals. Both fixed references are displayed for all six cases.

Table 6: All six WSP cases. p_T and the full-pool maximum use all 1,024 responses and are descriptive percentages. T_A is selected on the first 512 answers; Δ_B^T is its gain on the other 512 over fixed T, in points. The common references 0.1 and 0.9 are not validation-selected on these cases.

Case	$p_{0.1}$	$p_{0.9}$	Full-pool max (T)	T_A	$\Delta_B^{0.1}$	$\Delta_B^{0.9}$
WSP1	79.98	93.65	93.65 (0.9)	0.9	+13.48	+0.00
WSP2	40.23	31.25	44.63 (0.2)	0.2	+3.52	+10.55
WSP3	0.98	25.39	27.15 (1.1)	1.2	+19.14	-4.49
WSP4	0.00	0.10	0.39 (0.5)	0.5	+0.00	+0.00
WSP5	0.49	4.79	5.37 (0.8)	0.5	+2.34	-1.37
WSP6	0.00	0.29	0.59 (1.3)	1.1	+0.39	+0.00

every case, avoiding reference selection solely to favor a given example. Intervals in Figure 9 use 10,000 independent request-block bootstrap resamples per action. They are pointwise within-case intervals, not corrected claims across the temperature grid or uncertainty over the case population. The full dataset was available before this replay, so the split reduces within-analysis peak-selection bias but is not a fresh confirmatory experiment.

WSP1 and WSP3 retain positive replay gains over the source temperature; WSP2 retains a gain over 0.9. However, WSP3's selected temperature does not beat 0.9 on the replay pool, and the three error-containing prefixes offer little absolute accuracy. In WSP4, truncation reaches 99.90% at some temperatures; the corresponding maxima are 64.26% for WSP5 and 21.58% for WSP6. Low recovery is thus conditional on the recorded budget, not proof that no temperature or longer computation could repair the state. Neither these six selected cases nor their split-half gains measure how many root questions an automatic screen can identify. The larger MATH diagnostic and fresh screened-policy comparisons address separate parts of that question.

C Correct-count screen: data checks and sensitivity

We reconstruct the earlier screening report from per-prefix temperature histograms and source-response metadata, rather than copying plot constants. The original combined dataset has 350 prefixes. Two source records contain only one answer instead of eight; their raw counts had nevertheless been grouped as if they were counts out of eight. Removing them leaves 348 prefixes from 298 distinct MATH questions. These are earlier experimental partitions, not a newly untouched test set. Continuations and source probes are separate pools, but the prefix is taken from a source trajectory. Every response range uses eight candidate temperatures with 128 continuations each; these are not the 1,024-per-temperature WSP case studies.

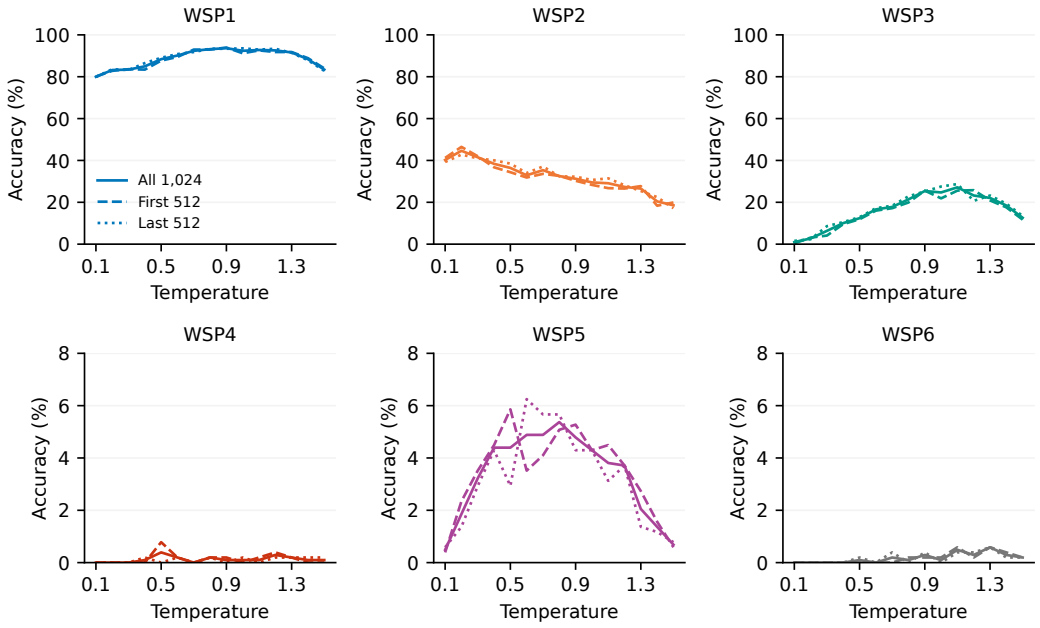


Figure 10: Full and split-pool curves for every WSP case. The bottom row uses a 0–8% scale to expose low-probability variation; the top row uses 0–100%. Temperature rankings need not match exactly between halves, even with 512 responses per temperature.

Table 7: Complete reconstructed screening diagnostic after excluding two non-eight-answer source records. Rates and mean ranges are prefix-weighted.

Source correct / 8	Prefixes	Large response (%)	Mean range (pp)
all	348	26.1	17.48
0 or 8	204	10.3	8.84
1–7	144	48.6	29.72
4–7	92	57.6	33.08
5–6	32	75.0	39.65
3–5	53	49.1	30.57
outside 3–5	295	22.0	15.13

The earlier large-response rule is the strict inequality $\max_T \hat{p}(T) - \min_T \hat{p}(T) > 0.25$. Before exclusions, we exactly recover the slide's group counts and rounded rates: 205 prefixes with 0/8 or 8/8 and 10.2% large response; 145 with 1–7/8 and 48.3%; 92 with 4–7/8 and 57.6%; and 32 with 5–6/8 and 75.0%. Table 7 reports the corrected counts and adds the current method's 3–5/8 rule. The narrower earlier intervals overlap and are not independent replications.

At the 25-point threshold, 91/348 prefixes have large response and 257/348 do not. Equal-question weighting gives 26.5% large response overall, 46.6% within 3–5/8, and 23.0% outside 3–5/8, preserving enrichment. This is threshold-dependent: at 10, 15, 20, and 30 points the overall large-response proportions are 50.0%, 40.8%, 35.1%, and 21.8%. The corresponding 3–5/8 proportions are 77.4%, 67.9%, 60.4%, and 43.4%. Thus "below the large-response threshold" must not be read as "temperature has no effect." The original best-versus-runner-up margin plot cannot establish invariance either: two temperatures can tie for best while other temperatures perform much worse.

The average signed 0.1-to-0.9 change is +0.61 points over all prefixes, despite a 17.48-point mean within-prefix range. On 3–5/8 prefixes the same uniform temperature change yields only +0.56 points, despite the enriched large-response rate. Screening finds sensitivity, not a guaranteed direction of improvement. Full-grid ranges are finite-sample, post-hoc estimates, and can include sampling noise; the disjoint-pool diagnostic addresses repeatability separately. This table estab-

Table 8: Observed and difficulty-preserving null response statistics. The large-range criterion is > 25 percentage points.

Statistic	Screened (53)	Outside (295)
Observed large-range proportion (%)	49.06	22.03
Null expected large-range proportion (%)	0.0424	0.0116
Observed mean range (pp)	30.57	15.13
Null expected mean range (pp)	9.19	5.31

lishes neither an exhaustive three-class taxonomy nor the exact fraction of near-zero-response root questions. The answer-assisted nature of the screen also remains explicit.

C.1 Enrichment beyond sampling noise

This retrospective analysis reuses the 348-prefix, 298-question MATH population above: 53 prefixes pass the 3–5/8 rule and 295 do not. Each prefix has 8×128 continuations, totaling 356,352 answers. No new model responses or training are used. All analyses retain the earlier strict range threshold $\widehat{\mathcal{R}} > 0.25$.

Difficulty-preserving conditional randomization. For each prefix, pool its 1,024 binary correctness outcomes, hold the total number correct fixed, and randomly allocate them without replacement to eight temperature groups of 128. We use 20,000 allocations per prefix. This null preserves average difficulty and its associated binomial sampling variance while removing temperature–correctness association. Table 8 contrasts observations with the expected null statistics; the null large-range rates are not p-values or estimates of the false-discovery fraction.

The observed large-range proportion gap is +27.02 points, with a 95% question-cluster bootstrap interval of $[10.94, 42.51]$. Each bootstrap draw resamples the 298 original questions and includes all their prefixes together; the estimand remains prefix-weighted. Equal-question weighting within each group gives 46.59% versus 23.03%. For each prefix we also compute the Monte Carlo tail probability of its observed range, using a plus-one correction (Phipson & Smyth, 2010). After Holm adjustment (Holm, 1979) over 348 tests, 31/53 screened prefixes (58.49%) and 102/295 outside prefixes (34.58%) reject temperature independence at 0.05. This tests whether an effect exists, not whether its true range exceeds 25 points.

As a threshold-free check, let $\hat{p}_j$ denote accuracy at temperature j. The binomial-noise-corrected between-temperature variance is

$$\widehat{V}_{\text{corrected}} = \frac{1}{8}\sum_{j=1}^{8}(\hat{p}_j - \bar{p})^2 - \frac{7}{8}\frac{1}{8}\sum_{j=1}^{8}\frac{\hat{p}_j(1-\hat{p}_j)}{127}. \tag{4}$$

Its group means are 0.01486 versus 0.00516, with gap interval $[0.00355, 0.01688]$. This correction is unbiased for the finite-grid variance under independent binomial sampling; it need not be nonnegative for each prefix.

Disjoint-half action selection and evaluation. We randomly split each temperature's 128 outcomes into 64/64 halves. On the first half, select the highest- and lowest-accuracy temperatures; on the second, evaluate their signed difference. Reverse the roles and average both directions, then average 2,000 such partitions per prefix. Binary histograms permit exact hypergeometric splitting without synthesizing new responses. Table 9 reports the original eight-action grid and two restrictions. Enrichment persists below $T = 1$, so it does not depend solely on high-temperature degradation. Intervals cluster original questions, not partitions; repeated splits do not increase the number of independent questions.

Selecting only the best action in one half and comparing it with fixed $T = 0.9$ in the other gives +10.56 points on screened prefixes versus +4.18 outside. The gap is +6.38 points with question-cluster interval $[1.95, 11.33]$. This uses 8×64 correctness-labeled answers per prefix for action selection; it is label-assisted opportunity evidence, not MLP prediction or superiority over the subgroup's best fixed temperature.

Table 9: Complete cross-evaluated high-minus-low temperature differences (percentage points). Intervals describe screened-minus-outside gaps.

Temperature range	Screened	Outside	Gap [95% CI]
0.1–1.5	26.45	12.58	+13.87 [6.23, 21.35]
0.1–1.1	19.19	8.42	+10.77 [4.46, 17.14]
0.1–0.9	17.00	7.40	+9.60 [3.80, 15.48]

Statistical scope. Finite-run evaluations require uncertainty-aware comparisons, as emphasized in reinforcement-learning evaluation by Agarwal et al. (2021). Our intervals concern question sampling, not variation across training runs. The randomization assumes exchangeable outcomes within a prefix under the null; the histograms cannot verify potential generation-batch correlations. Question- cluster intervals account for repeated prefixes but are conditional on the stored answer pools and historical screening rule. Except for the Holm-adjusted prefix tests, sensitivity intervals are pointwise and not jointly multiplicity- adjusted. All checks are retrospective rather than a new untouched confirmation experiment. They support genuine response enrichment under the sampling model, not independence from difficulty, exact true-range prevalence, or a causal validation of root-prompt screening. Bootstrap uses 5,000 draws with seed 20260925.

D Selector development and evaluation protocol

All selector experiments use frozen Qwen3-8B (Yang et al., 2025) with a common mathematical reasoning prompt. MATH supplies the training and calibration data; GSM8K (Cobbe et al., 2021) and OlympiadBench receive the exported MLP without refitting. This section consolidates representation selection, final-model training, and evaluation. The final policy's results appear in Appendix E; other selector experiments are kept separate in Appendix G.

D.1 Representation development

Before target-domain evaluation, a MATH-only probe compared 19 cached feature groups under the same small MLP. It used 800 prefixes from 400 fit questions, grouped five-fold out-of-fold (OOF) predictions, and an outer 50-question development split. The label was the tie-aware empirical best vote-margin temperature over eight actions. Table 10 reports representative candidates; lower cross-entropy is better. "Hit" counts a prediction as correct if it belongs to the empirical set of tied best actions. The constant predictor uses no state feature. The raw layer-36 64-token mean plus length was selected by OOF before inspecting its outer-development result.

Table 10: Representative MATH-only feature-development results. These losses select a representation and do not measure answer accuracy. The selected raw layer-36 mean plus length is best by grouped OOF CE and outer-development CE.

Representation	Input dim.	OOF CE	Outer-dev. CE	Outer-dev. hit (%)
Constant predictor	0	1.9529	1.9517	52
Layer-36 last token	4,096	1.9454	1.9747	52
Length only	4	1.9143	1.9569	50
Layer-18 trailing-64 mean	4,096	1.9079	1.9554	54
Layer-36 trailing-64 mean	4,096	1.9106	1.9179	60
Four-layer trailing-64 means	16,384	1.9187	1.9469	54
Probability summaries	140	1.9819	1.9442	60
PCA32 length + layer-36 mean	32	1.9193	2.0068	–
Raw length + layer-36 mean (selected)	4,100	**1.9044**	**1.9115**	**64**

The cached probe has four prefix-length coordinates, whereas the deployable prompt-only representation drops the two generated-prefix coordinates that are undefined before generation, leaving 4,098 inputs. The probe also shows why we do not use PCA: for the selected feature family, removing PCA changes outer development CE from 2.0068 to 1.9115. This is a comparison between com-

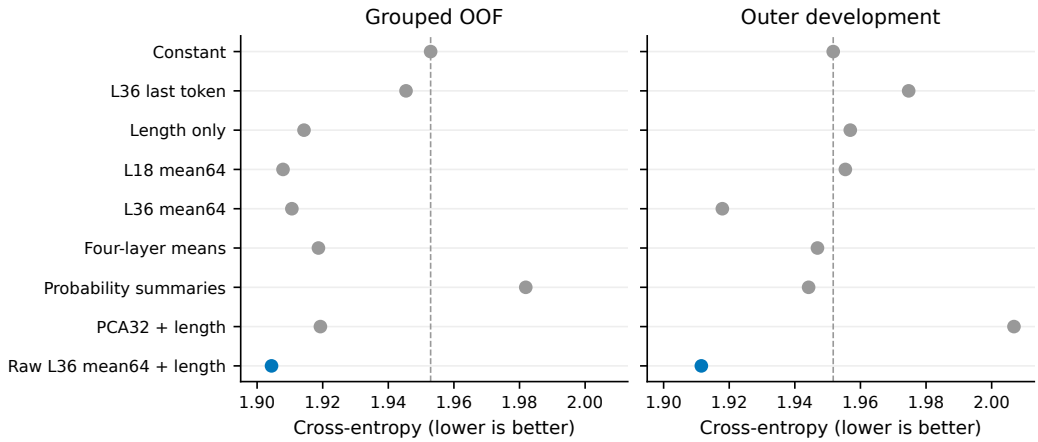


Figure 11: Feature-development point estimates from Table 10. All nine representative configurations are shown in the table's order. Dashed lines mark the no-state constant predictor. The blue point identifies the raw layer-36 mean plus length selected by grouped OOF loss; the outer split is reported separately. These are vote-margin temperature-label losses, not answer accuracy or a controlled estimate of PCA information loss.

plete modeling pipelines with different parameter counts, not a pure causal estimate of information destroyed by PCA.

D.2 Final prompt-only MLP: labels and optimization

The 919 MATH candidates were themselves historically prescreened. A new eight-answer screen at $T = 0.1$ retains 395 questions, partitioned by question into 280 fit, 70 calibration, and 45 evaluation roles. Fit and calibration questions each supply 16 complete-answer correctness labels at each of eight temperatures, totaling 44,800 labels. Of these, 1,088 are regraded, prompt-matched saved responses from 17 fit questions; 43,712 are newly collected. All calibration labels are fresh. No evaluation response or screen answer is a training target.

For question q, define

$$\hat{p}(q,T) = \frac{1}{16}\sum_{j=1}^{16} R_{q,T,j}, \qquad a(q,T) = \hat{p}(q,T) - \frac{1}{8}\sum_{t\in\mathcal{A}} \hat{p}(q,t). \tag{5}$$

Let $f_\theta(z(q)) \in \mathbb{R}^8$ be the raw network output and $\widehat{A}_\theta(q,T) = f_{\theta,T}(z(q)) - \frac{1}{8}\sum_{t\in\mathcal{A}} f_{\theta,t}(z(q))$. The data loss is the equal-question, equal-action mean squared error

$$\mathcal{L}(\theta) = \frac{1}{280\cdot 8}\sum_{q\in\mathcal{D}_{\text{fit}}}\sum_{T\in\mathcal{A}} \big(\widehat{A}_\theta(q,T) - a(q,T)\big)^2. \tag{6}$$

This is regression on centered individual correctness, not classification, cross-entropy, or direct optimization of SC@8.

Features concatenate $\log(1+n)$, $n/8192$, and the 4,096-dimensional final normalized layer-36 hidden-state mean over the last $\min(n, 64)$ prompt tokens. The saved Qwen chat template uses thinking disabled and a step-by-step system instruction ending in a boxed answer. Each coordinate is standardized using fit-only means and population standard deviations; constant coordinates use unit scale. No PCA or feature selection is fitted on calibration/evaluation data. The network is Linear(4098,64)–GELU–Linear(64,32)–GELU–Linear(32,8), with biases, 264,680 trainable parameters, no dropout, and no output softmax. The last layer starts with zero weights and the fit-average target as its bias, so epoch 0 is a constant action-profile predictor.

Optimization uses float32, seed 42, and full-batch AdamW on all 280 fit questions: learning rate 10^{-3}, weight decay 10^{-2}, betas $(0.9, 0.999)$, epsilon 10^{-8}, and at most 200 epochs, without a

learning-rate schedule. The 70 calibration questions select the checkpoint by the same centered MSE, requiring improvement greater than 10^{-7}; patience is 25 epochs. Training stops at epoch 30 and restores epoch 5 (fit MSE 0.0162024, calibration MSE 0.0177379). The same calibration set subsequently selects the model class and gate by individual accuracy. This reuse is development, not independent confirmation. There is no post-selection refit, target-domain training, reinforcement learning, or language-model weight update. The exported gate compares the best predicted action with $T_0 = 0.9$ and switches only for a strictly greater than 0.05 predicted difference; action ties prefer T_0, then ascending temperature. The threshold is not a confidence level or a guarantee of a five-point realized gain.

D.3 Prompt-only policy calibration

The final advantage model is refit on prompt-only features and correctness targets, rather than reusing the probe's vote-margin classifier. Table 11 gives all seven prespecified calibration candidates on 70 MATH questions. Figure 11 summarizes the separate representation probe. Random forests follow Breiman (2001). The gate reverts to fixed $T = 0.9$ unless predicted advantage exceeds the threshold. MLP plus a 0.05 gate has the largest point estimate, although its +0.09-point calibration edge is too small to count as independent efficacy evidence. Its value is procedural: model class and gate are fixed without looking at either target dataset.

Table 11: MATH prompt-only policy calibration used to freeze the deployed selector.

Model	Gate	Individual accuracy (%)	Switched questions
Fixed $T = 0.9$	–	58.04	0
Random forest	0	55.98	70
Random forest	0.02	55.98	39
Random forest	0.05	58.04	0
MLP	0	55.89	65
MLP	0.02	56.52	33
MLP (selected)	0.05	**58.13**	11

D.4 Data and response accounting

Table 12 lists the candidate and retained populations. Screen answers are separate from downstream evaluation answers. The exported model actions are replayed from saved features before target-domain answer scoring; each action is frozen before response generation.

Table 12: Question and response accounting for the prompt-only evaluation. "Retained" is defined by the answer-assisted 3–5/8 screen.

Dataset	Candidate frame	Retained	Rejected	Matched control	Downstream response budget
MATH	919	395	524	–	fit/cal.: 16 per temperature; eval.: 32 per policy
GSM8K	512	30	482	30	8 per policy and question
OlympiadBench	512	56	456	56	8 per policy and question

The 395 retained MATH questions are frozen into 280 fit, 70 calibration, and 45 evaluation roles. The fit/calibration label tensor contains 44,800 correctness labels over eight temperatures. Every dataset additionally uses eight answers per candidate for screening; those answers are disjoint from downstream target evaluation pools. Equal response counts and token caps do not imply equal realized token expenditure.

D.5 Evaluation protocol and random rejected controls

Target-domain screening and fixed baselines. For each target dataset, the screen in Equation 3 is computed from eight independent answers at $T = 0.1$, generated from the question prompt and frozen before selector evaluation. It retains 30 of 512 GSM8K questions and 56 of 512 OlympiadBench questions. For each target dataset, a separate, unscreened 64-question development split

selects the fixed baseline from the eight temperatures in $\mathcal{A}$. We draw eight complete answers per question–temperature pair and maximize pooled *individual-answer accuracy*, breaking ties by grid order. This gives $T^* = 1.1$ for GSM8K and $T^* = 0.9$ for OlympiadBench. Neither screened evaluation questions nor their fresh answers enter this choice. On MATH, the source-domain fixed $T^* = 0.9$ is selected by the same criterion on its separate 70-question calibration split. The selector never consumes screen answers; it only uses the rendered question prompt.

Prompt-only selection and evaluation. For every retained question, we compute and freeze the MLP action before the evaluated answers are generated. We then collect eight fresh complete answers at the selected temperature and at T^*. If the two actions coincide, they share the same eight completions; otherwise they use disjoint deterministic seed namespaces. The original prompt-only transfer protocol names equal-question individual accuracy as primary and SC@8 as secondary. The subsequent rejected-control protocol instead calls SC@8 its primary task metric. We preserve this history rather than claim a single prespecified endpoint hierarchy across all stages. Throughout this paper, individual accuracy is the original transfer endpoint and the selector's training/calibration objective; SC@8 is a separate voting endpoint. SC@8 counts a correct unique normalized-answer mode, with ties wrong. We report both and use a 5,000-draw question bootstrap interval for paired individual-accuracy differences. Differences are computed before rounding. These fresh answers do not overlap with screen samples, MATH labels, or development samples. Importantly, the fixed baselines were selected for individual accuracy, not for SC@8; an SC@8 difference against T^* is therefore not a comparison with an SC@8-optimized fixed baseline. The completed-grid analysis separately compares SC@8 with development-selected T^*_{SC}; its zero GSM8K and negative OlympiadBench point estimates govern claims of SC@8-matched improvement. This does not change the individual-accuracy contrasts.

Random rejected controls. We pre-freeze one same-size random control per target dataset from the questions rejected by the 3–5/8 screen. GSM8K samples 30 without replacement from 482 rejected questions using seed 202609211; OlympiadBench samples 56 from 456 rejected questions using seed 202609212. Controls use the same frozen selector, fixed baseline, answer budget, prompt-only policy, and scorer as the screened sets. They are matched in size and evaluation procedure, not in difficulty. Table 2 reports both groups' gains.

Generation and scoring. Each target-policy evaluation uses eight complete answers per question and an 8,192-output-token cap per answer. Semantic answer scoring is replayed independently after generation. Equal response counts and token caps do not imply equal realized token expenditure. Completing the eight-temperature grid retains existing evaluation cells and generates only missing cells; no selector is refitted and no screen answers are reused. Appendix E.2 gives the complete curves and their response accounting.

Statistical scope. The results use one frozen rollout block and one rejected-control draw per target. Question-bootstrap intervals do not include training or architecture selection variability, repeated screening, or between-rollout variability. Root-policy gain-gap intervals are unavailable. We report directional evidence, not statistically settled cross-domain superiority. Attributing gains to learned question–temperature matching would additionally require action-frequency-matched controls; the modal-policy decomposition alone does not establish that mechanism.

D.6 Experimental populations and dataset construction

The study combines complementary experiments rather than one pooled benchmark. Table 13 distinguishes their units and evidential roles. In particular, varying the continuation temperature at a fixed prefix is not the same intervention as choosing a temperature before any answer is generated. The multi-model sweeps are descriptive response studies, not replications of the Qwen3-8B selector. Prefix-feature development informs the representation choice; the final prompt-only model is trained anew on correctness-derived targets.

MATH candidate construction. The 919 candidates comprise 811 training-role and 108 validation-role questions from an existing MATH pool prescreened for 3–5 correct answers out of eight at $T = 0.1$. These are project-level roles, not the original MATH benchmark splits.

Table 13: Experimental boundaries. Counts denote the named study populations, not independent replications of the final selector.

Study	Population and sampling unit	What it establishes
Multi-model MATH sweep	Three Qwen3 models; five difficulty strata; one logged aggregate per temperature	Aggregate temperature response
High-sampling AIME cases	Qwen3-8B; six curated fixed prefixes; 1,024 continuations per action	Local response patterns
MATH screen diagnostic	348 fixed prefixes from 298 questions; 128 continuations per action	Screening enrichment and disjoint-pool opportunity
Representation probe	800 prefixes from 400 fit questions; 50 outer-development questions	Choice of feature family, not prompt-only efficacy
Final MATH MLP	280/70/45 prompt-level fit/calibration/evaluation questions	Training and source-domain policy evaluation
Target-domain transfer	30 GSM8K and 56 OlympiadBench retained questions, plus same-size rejected controls	Frozen prompt-only policy comparisons

The earlier split was stratified by subject and difficulty with seed 42; questions used by the earlier PVM training population were assigned to its training role. The separately reserved 162-question project test split is not used here. A new eight-answer screen yields 350 retained training-role questions and 45 retained validation-role questions. Sort the 350 by the SHA-256 digest of `root-learning-v1|42|` followed by question ID: the first 70 form calibration and the remaining 280 form fit. The 45 retained validation-role questions form evaluation. These roles are disjoint for final-model fitting and evaluation; the candidate pool has a development history and is not claimed to be an untouched sample from the full MATH benchmark.

Target datasets and deterministic selection. GSM8K uses 64 development questions from its official training split and 512 candidate evaluation questions from its official test split. OlympiadBench (He et al., 2024) uses the Hugging Face configuration `OE_TO_maths_en_COMP`, whose repository split is named `train`. We retain English, text-only, open-ended mathematics questions with exactly one reference answer; multi-answer items, image-bearing items, and empty questions or answers are excluded by schema, not model performance. Context text, when present, is prepended to the question. After normalization by whitespace removal and case folding, exact duplicate questions and exact matches to the original 69-question prefix-selector development cohort are removed. This exact-text check does not certify absence of semantic duplicates or pretraining overlap.

Eligible questions are ordered by SHA-256 of `screen-transfer-v1|20260915|` followed by question ID. GSM8K takes the first 64 eligible training and 512 eligible test questions; OlympiadBench takes the first 64 and next 512 from its shared eligible pool. The pinned source revisions are `3101c7d50724` for the official GSM8K repository and `91184b52131e` for OlympiadBench; the full revision identifiers are recorded in the supplementary preparation code. Screening, prompt-only evaluation, and prefix controls reuse these frozen question identities but use separate generated response pools.

D.7 Prompt, decoding, and scoring contract

The final selector experiments use `Qwen/Qwen3-8B` and its chat template with `enable_thinking=False` and `add_generation_prompt=True`. The user message contains only the problem, including dataset-provided context where applicable. The system instruction is:

> You are a math reasoning assistant.
> Formatting rules:
> - Solve the problem step by step.
> - Each step must be written as a separate paragraph.
> - Separate every step with exactly two newline characters.
> - Do not use numbering, bullets, or any markers at the start of a step.

- Each step must be a complete sentence that describes one reasoning move.
- Inline LaTeX expressions should use `$...$`.
- The final paragraph must include the final boxed answer written as `\boxed{}`.
- Do not include any explanations, headers, or summaries outside the steps.
- The response must end immediately after the final paragraph containing `\boxed{}`.

The exact serialized instruction is also in the supplementary dataset config. Sampling uses top-$p = 1$, unrestricted top-k (`-1` in vLLM), an 8,192-generated-token cap, and a 10,240-token context limit. For prefix controls, the prefix consumes part of the same 8,192-token response budget. The final prompt-only comparisons generate no prefix before selecting an action.

Answer extraction uses the last closed box, with gold-independent answer normalization followed by bounded semantic mathematical equivalence checking. Missing answers, verification failures, and truncated completions count as incorrect. SC@8 requires a unique highest-vote normalized answer that is correct; it does not require an absolute majority. Tied modes and a winning missing-answer bucket count as failures. MATH evaluation uses 32 answers per question and policy; its four-block SC@8 summary differs from each target task's single eight-answer block. The archived request-seed functions separate screen, training-label, development, and evaluation namespaces; equal policy actions share a response cell as specified in Appendix D.5.

The recorded environment uses Torch 2.10.0, Transformers 4.57.6, vLLM 0.18.1, scikit-learn 1.7.2, math-verify 0.9.0, NumPy 2.2.6, and joblib 1.5.2. The supplement records language-model weight hashes and includes the frozen MLP checkpoint, coordinate transform, and selector settings. These settings describe the final selector experiments; the multi-model sweeps and AIME continuation cases have their separately documented protocols.

D.8 Anonymous supplementary material and numerical replay

The anonymous code supplement includes screening and fitting code, frozen selector artifacts, question IDs and role assignments, derived evaluation outcomes, a minimal MATH normalized-answer export, protocol documents, and a SHA-256 manifest. From its root, running `python3 reproduce_tables.py` requires only the Python standard library. It verifies the screen and role counts, reconstructs MATH individual accuracy and the four SC@8 blocks, and recomputes the screened/rejected target comparisons and all eight fixed-temperature points on the four target cohorts.

This replay uses stored outcomes: it neither regrades generated text nor re-estimates confidence intervals. The release-level and temperature-curve tests check the supported accounting and voting logic. Public benchmark problems and Qwen3-8B weights must be obtained separately for generation; the archive does not distribute raw reasoning traces or all training labels. Historical execution helpers retain their original provenance contracts and are not a turnkey end-to-end pipeline. The accompanying paper-source package separately contains descriptive aggregates and rendering code, including the screening evidence table and the selected W&B accuracy curves.

E Complete results for the final prompt-only selector

This section reports the final frozen MLP and its fixed-temperature comparisons. The other selectors and prefix-based populations in Appendix G are not replications of this policy.

E.1 MATH source-domain evaluation

The no-PCA prompt MLP is evaluated before answer generation on 45 MATH questions with 32 completions per policy and ties fixed $T = 0.9$ at 54.17% individual accuracy (two wins, two losses, and 41 ties by question). Historical prefix-based results use different policies and populations and are reported separately in Appendix G.1.

For completeness, we also partition each prompt-only policy's 32 saved completions into four non-overlapping groups of eight and apply the same unique-plurality SC@8 rule as in the target experiments. Across 180 question-blocks, fixed $T = 0.9$ scores 120/180 (66.67%) and the MLP scores 123/180 (68.33%), a +1.67-point descriptive difference. The first block ties, while each of the other

three blocks gains one vote-correct outcome. All three changes occur on the same switched question, so the 180 blocks are not treated as independent questions and this post-hoc calculation is not presented as a robust source-domain effect.

E.2 Complete target-domain temperature comparisons

The sweep covers $\{0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5\}$ on the same screened and random rejected-control questions as Tables 3 and 2. It contains 172 distinct questions and 11,008 answers: 1,936 existing evaluation answers and 9,072 newly generated answers. Original complete question–temperature cells are retained unchanged; missing cells receive eight fresh answers in a separate, fixed seed namespace. When a selector action equals a fixed temperature, both policies use that cell's identical response pool. The curves therefore supplement the original comparison rather than independently replicate it. In particular, a constant selector must exactly coincide with its corresponding fixed-temperature point.

All cells use the original frozen model, chat template, answer normalization, and correctness checker. Truncated generations count as failures. SC@8 uses the unique most frequent normalized answer; tied modes count as failures. Each question has equal weight. The 95% intervals resample questions with replacement 5,000 times, with shared indices for paired differences. They are pointwise intervals conditional on the saved response pools, not simultaneous bands over eight temperatures and not estimates of between-run uncertainty. In particular, a zero-width interval at 100% does not imply certainty about population performance. The empirical curve maximum is selected after seeing evaluation outcomes and is reported only as a descriptive reference. These curves concern the named subsets, not the full GSM8K or OlympiadBench benchmarks. Screening and development responses are excluded from every evaluation point; question identities, prompt hashes, vote slots, and original headline metrics are checked before export.

Fixed baselines selected for SC@8. Figure 12 reports the full screened SC@8 sweeps for both target tasks. Relative to the original fixed temperatures selected for individual accuracy, the SC@8 gains are +6.67 points on GSM8K and +3.57 on OlympiadBench; the corresponding screen-minus-control gain gaps are +6.67 and +5.36 points. The rejected controls are easier, with both policies attaining 100% SC@8 on the GSM8K control. The metric-matched comparisons below instead select each fixed baseline for SC@8 itself.

On the 64-question GSM8K development split, fixed $T = 0.9$, 1.1, and 1.3 each obtain 59/64 correct votes. Ascending-grid tie breaking chooses $T = 0.9$, which is exactly the frozen selector's action on every screened and rejected GSM8K question. Their SC@8 differences are therefore zero using shared pools. On OlympiadBench development questions, fixed $T = 0.3$ has the highest SC@8 score (40/64, versus 38/64 at $T = 0.9$). The completed grid includes this comparator, whereas the main table retains the original baselines selected for individual-answer accuracy. Neither baseline uses evaluation outcomes for its choice; the highest point on the evaluation curve is a different, post-hoc reference. On the screened OlympiadBench questions, this SC@8-selected fixed $T = 0.3$ achieves 40/56 correct votes (71.43%), versus 34/56 (60.71%) for the selector. The difference is -10.71 points, with a paired question-bootstrap 95% interval of $[-25.00, +3.57]$. This reverses the direction of the +3.57-point comparison against the individual-accuracy-selected $T = 0.9$; the interval does not establish population-level inferiority either. On the matched OlympiadBench rejected control, fixed $T = 0.3$ obtains 62.50% SC@8 versus 66.07% for the selector (+3.57 points; paired 95% interval $[0.00, +8.93]$). Thus the screen-minus-control SC@8 gain gap changes from +5.36 points under fixed $T = 0.9$ to -14.29 points under fixed $T = 0.3$. The GSM8K gap becomes zero because the selector equals its SC@8-selected fixed policy in both populations. These are point-estimate comparisons, not a separately tested gain-gap hypothesis.

E.3 Fixed temperatures selected on screened development questions

To examine population alignment of the comparator, we retrospectively apply the same 3–5/8 screen to each task's separate 64-question development set. Only two GSM8K and nine OlympiadBench questions pass. Development and evaluation question identities are disjoint; neither selector actions nor evaluation populations are changed.

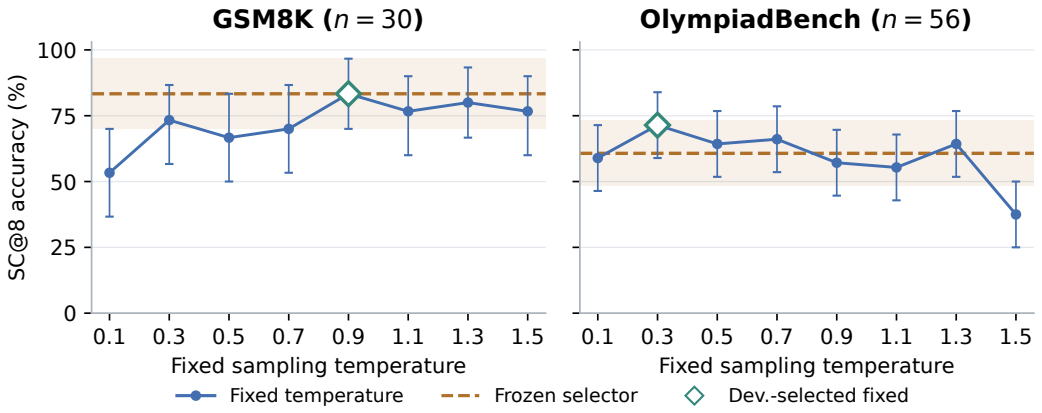


Figure 12: Complete SC@8 curves on screened GSM8K and OlympiadBench questions. Dashed lines show the frozen selector; diamonds mark fixed temperatures selected on separate development questions to maximize SC@8. Each question–temperature cell has eight answers, with tied votes counted as failures. Error bars and shaded bands are pointwise 95% question-bootstrap intervals conditional on saved response pools. Identical selector and fixed actions share responses. Both target tasks are shown under the same protocol.

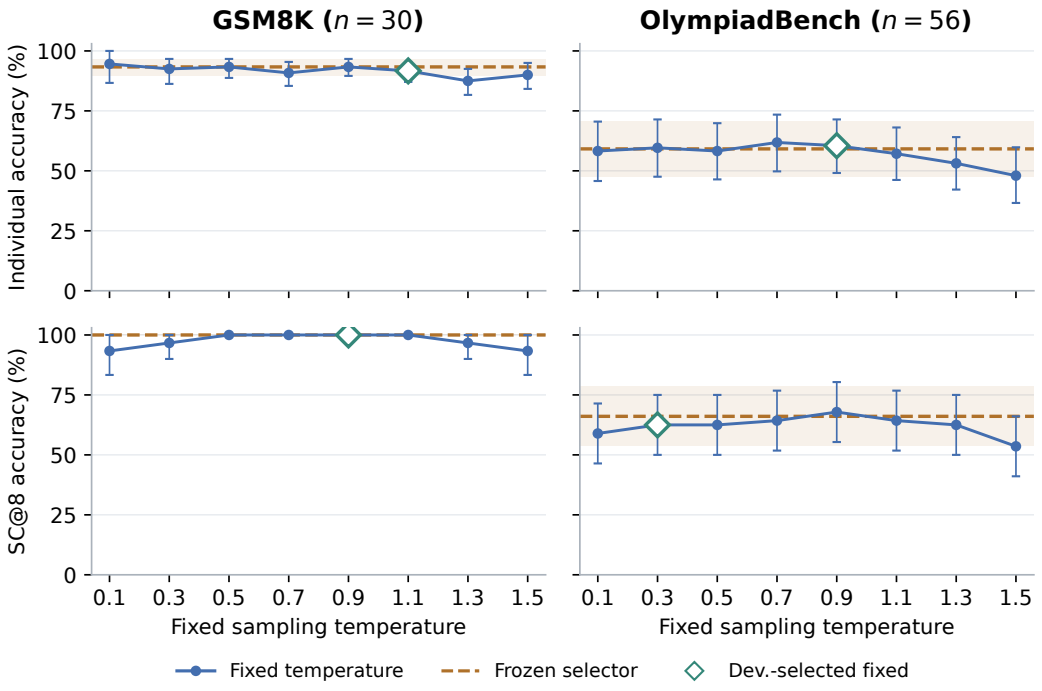


Figure 13: Complete fixed-temperature curves on the matched random rejected controls. Dashed lines are the frozen selector; diamonds mark the fixed temperature selected on disjoint development questions for the plotted endpoint. Error bars and shaded selector bands are pointwise 95% question-bootstrap intervals conditional on one response pool. Each policy uses eight answers per question; neither panel represents a full benchmark.

Each development question has only one eight-answer pool at $T = 0.1$. Once that pool defines screening, it cannot independently score the same action. We therefore select the fixed temperature from the other seven actions, $\{0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5\}$, maximizing the relevant endpoint and

Table 14: Complete sweep on screened questions: accuracy in percent, with eight completions per question and temperature.

	GSM8K		OlympiadBench	
Temperature / policy	Individual	SC@8	Individual	SC@8
0.1	57.50	53.33	45.31	58.93
0.3	59.58	73.33	48.66	71.43
0.5	57.50	66.67	46.88	64.29
0.7	61.67	70.00	48.66	66.07
0.9	65.42	83.33	46.65	57.14
1.1	58.33	76.67	39.73	55.36
1.3	63.33	80.00	39.06	64.29
1.5	59.58	76.67	28.57	37.50
Frozen selector	65.42	83.33	48.21	60.71

Table 15: Complete sweep on matched random rejected controls: accuracy in percent.

	GSM8K		OlympiadBench	
Temperature / policy	Individual	SC@8	Individual	SC@8
0.1	94.58	93.33	58.26	58.93
0.3	92.50	96.67	59.60	62.50
0.5	93.33	100.00	58.26	62.50
0.7	90.83	100.00	61.83	64.29
0.9	93.33	100.00	60.49	67.86
1.1	91.67	100.00	57.14	64.29
1.3	87.50	96.67	53.12	62.50
1.5	90.00	93.33	47.99	53.57
Frozen selector	93.33	100.00	59.15	66.07

breaking ties in ascending grid order. Including the reused $T = 0.1$ point happens to leave the selected actions unchanged, but does not resolve its dependence. This is a seven-action sensitivity check, not a fully independent eight-action screened-development experiment.

Table 16 shows that GSM8K's fixed action changes from 1.1 to 1.3, reducing the selector gain from +7.08 to +2.08 points. OlympiadBench retains fixed 0.9 and +1.56 points. Positive point estimates remain, but the small screened development sets limit the stability of this comparison. In 10,000 conditional question-bootstrap draws from the screened development sets, GSM8K selects 1.3 in 75.06% and 1.1 in 24.94%; leave-one-out selects each once. OlympiadBench selects 0.9 in 66.97% of draws and all nine leave-one-out fits. These frequencies do not include variability from repeating the screen.

On GSM8K's two screened development questions, $T = 0.7, 0.9, 1.1, 1.3$ all score 2/2 correct votes. Their screened-evaluation selector-minus-fixed differences are respectively +13.33, 0.00, +6.67, and +3.33 points. The +13.33-point interval in Table 17 conditions on choosing 0.7; it does not account for developmental selection uncertainty and cannot establish superiority over all tied choices. In particular, the selector equals fixed 0.9 on every GSM8K evaluation question. On OlympiadBench, screened development still uniquely selects 0.3 for SC@8, and the selector difference remains -10.71 points. The audit uses a different bootstrap seed from the original comparison, so interval endpoints can differ slightly without changing point estimates.

MATH already uses screened development: all 70 calibration and 45 evaluation questions meet the original independently sampled 3–5/8 rule, with disjoint question roles. The original fixed 0.9 is selected on those calibration questions. Its evaluation individual accuracy and the MLP's are both 54.17%, with paired difference interval $[-1.25, +1.25]$ points from 32 answers per question and policy.

Table 16: Individual accuracy versus a screened-development-selected fixed temperature. Δ is MLP minus fixed in percentage points on the unchanged 30-question GSM8K and 56-question OlympiadBench screened evaluation subsets. Intervals condition on the selected comparator.

Dataset	Dev. n	Fixed T	Fixed (%)	MLP (%)	Δ	95% CI
GSM8K	2	1.3	63.33	65.42	+2.08	[-4.17, +8.33]
OlympiadBench	9	0.9	46.65	48.21	+1.56	[-0.67, +4.46]

Table 17: SC@8 under screened-development selection. GSM8K's displayed fixed $T = 0.7$ results from the ascending-grid tie rule among four development-optimal actions; their complete evaluation contrasts are reported in the text.

Dataset	Dev. n	Fixed T	Fixed (%)	MLP (%)	Δ	95% CI
GSM8K	2	0.7	70.00	83.33	+13.33	[+3.33, +26.67]
OlympiadBench	9	0.3	71.43	60.71	-10.71	[-26.79, +3.57]

This sensitivity analysis reuses existing outcomes, performs no training or new generation, and preserves the original main-table comparisons. The portable aggregate records all tied actions and conditional selection frequencies. An accompanying cache check recomputes all 1,920 screened GSM8K and 3,584 screened OlympiadBench question–temperature answers from saved generation journals and matches the per-question summaries. It retains the original correctness labels; it is not independent semantic regrading.

E.4 Where do temperature-selection gains come from?

Gains can arise from adopting a better common temperature for screened questions or from assigning different temperatures to individual questions. Compare the development-selected constant T^*, selector π, and constant modal temperature T_{mode} on the same question set P. The mode is the selector's most frequent action, not an accuracy-selected optimum. Writing $A_P(\rho)$ for equal-question accuracy under policy ρ, adding and subtracting $A_P(T_{\mathrm{mode}})$ gives the accounting identity

$$A_P(\pi) - A_P(T^*) = \underbrace{A_P(T_{\mathrm{mode}}) - A_P(T^*)}_{\text{common-temperature gain}} + \underbrace{A_P(\pi) - A_P(T_{\mathrm{mode}})}_{\text{gain over the modal policy}}. \tag{7}$$

Table 18 reports both terms for each endpoint. The mode is a post-hoc reference; deviation from it is not a causal estimate of learned matching.

On GSM8K the frozen MLP selects $T = 0.9$ for all 30 retained questions, so the full gain comes from a shared temperature change on the screened subgroup. The same constant policy gains +1.67 individual-accuracy points on the rejected control, versus +7.08 on the screened population (Table 2).

OlympiadBench has the complementary profile. Because both T^* and the modal selector action equal 0.9, the full screened gain is assigned to five prompt-specific deviations: the MLP selects $T = 0.5$ on three questions and $T = 0.3$ on two, producing three wins, one loss, and one tie. The rejected control also has exactly five changed actions, but they produce zero wins, two losses, and three ties. This contrast identifies where the observed prompt-specific changes contribute positively, complementing the common-temperature profile on GSM8K.

F Reliability of temperature-advantage supervision

Fixed split-half construction. For every MATH question and temperature, the 16 saved correctness labels are partitioned without search into slots 0–7 (half A) and 8–15 (half B). Each statistic is computed independently in the two halves, and Table 19 reports their Pearson correlation. "Question

Table 18: Gain relative to a common-temperature policy (percentage points). "Common" is modal temperature minus fixed baseline; "Beyond mode" is selector minus modal temperature. MATH SC@8 is a post-hoc four-block calculation.

Dataset	Endpoint	T^*	T_{mode}	Common	Beyond mode	Total
MATH	Individual	0.9	0.9	0.00	0.00	0.00
MATH	SC@8 (post-hoc)	0.9	0.9	0.00	+1.67	+1.67
GSM8K	Individual	1.1	0.9	+7.08	0.00	+7.08
GSM8K	SC@8	1.1	0.9	+6.67	0.00	+6.67
OlympiadBench	Individual	0.9	0.9	0.00	+1.56	+1.56
OlympiadBench	SC@8	0.9	0.9	0.00	+3.57	+3.57

mean" first averages over temperature. "Centered action" subtracts each question's mean. "Question $\times$ temperature" additionally subtracts the global mean for that temperature and restores the grand mean. The last row computes the per-question difference between $T = 0.5$ and $T = 0.9$. Thus the two columns are separate within-population repeatability estimates; they are not a correlation between fit and calibration sets and do not measure MLP accuracy.

Table 19: Split-half reliability of MATH correctness-derived targets. The fit column excludes 17 questions with any legacy response, leaving 263 fresh-label questions; calibration contains 70 questions.

Target computed independently in A/B	Fresh fit r	Calibration r
Question mean accuracy	0.9097	0.9373
Question-centered action accuracy	0.2864	0.2186
Question $\times$ temperature interaction	0.2466	0.1931
$T = 0.5$ minus $T = 0.9$ contrast	0.0623	0.0858

These values reveal a difficulty–advantage identifiability gap: overall difficulty is repeatable under resampling, while the action-dependent residual is much noisier at eight responses per temperature. Low reliability complicates learning from these empirical labels, but does not prove that the latent advantage is absent or that no alternative representation can predict it.

On the same 263 fresh-label fit questions, choosing a temperature from eight responses per action and evaluating it on the other eight gives a +1.95-point individual-accuracy gain, averaged over both split directions. Selecting and evaluating on the same half instead gives +20.53 points. This diagnostic uses correctness labels rather than a predictor of temperature from prompts.

Fixed-budget prefix diagnostics. The curves in Figure 14 use the recorded training population of 400 questions with two low-temperature prefixes each and 128 saved continuations per temperature. Zero-based even slots form pool A and odd slots form pool B. At each $n \in \{8, 16, 32, 64\}$, accuracy correlations use the first n slots in each pool, with no split search or new generation. Replace "question" by "prefix" in the centering definitions above; all 800 prefixes are retained. Each question contributes exactly two prefixes, so pooled state weights are balanced by question. Temperature cells remain dependent.

Table 20: Accuracy-label repeatability in the same prefix population. Point estimates at four nested response budgets.

Responses per temperature per half	8	16	32	64
Mean accuracy	0.9959	0.9979	0.9990	0.9995
Prefix-centered accuracy	0.4347	0.6191	0.7629	0.8658
Prefix $\times$ temperature interaction	0.4268	0.6125	0.7575	0.8622
$T = 0.5$ minus $T = 0.9$	0.1286	0.2973	0.4734	0.6301

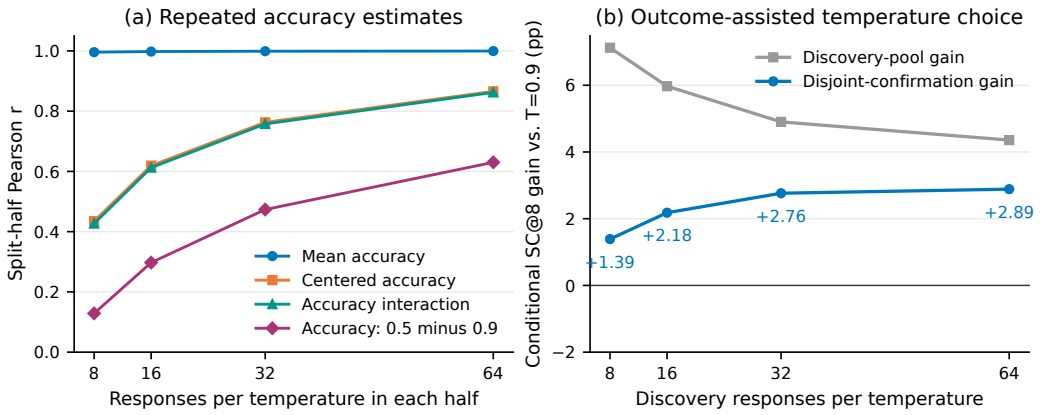


Figure 14: Complete sample-budget analysis for the recorded 400 MATH questions (800 prefixes). (a) Split-half accuracy correlations. (b) Outcome-assisted selection versus development-fixed $T = 0.9$, evaluated in the discovery pool or the fixed disjoint 64-response confirmation pool using exact finite-pool SC@8. All four nested budgets are shown as point estimates, not independent runs or learned-selector results.

For the utility curve, estimate unique-plurality SC@8 exactly by averaging over all eight-response subsets sampled without replacement from a saved pool. A tied or invalid winning answer counts as incorrect. Choose the temperature with greatest discovery utility; ties prefer $T = 0.9$, then ascending grid order. The comparator remains $T = 0.9$, selected on the separate 50-question development split. Confirm each choice on all 64 odd-slot continuations, independently of discovery budget. Average the two prefix utilities within each question, then across 400 questions. Discovery-pool gains are +7.12/+5.97/+4.90/+4.36 points, versus disjoint-pool gains of +1.39/+2.18/+2.76/+2.89 points. No fitted predictor chooses these actions. The exact finite-pool SC@8 expectation is not a new sampled SC@8 run or a guarantee for future generations. At 64 responses per half, the prefix-centered SC@8 profiles have split-half correlation 0.839, distinct from the doubly centered individual-accuracy interaction correlation of 0.862.

Observable response versus learned control. The positive label-based diagnostic does not establish learned selection. A separate prespecified single-seed experiment fitted capability-only, SC@8-advantage-only, and joint 64-width MLPs on 320 prefix training questions, using 80 for early stopping. On 250 evaluation questions (300 prefixes), capability-only NLL improves from the training-prior baseline 0.69048 to 0.65928. Advantage-only and joint centered-utility MSE are 0.01647 and 0.01708, versus 0.01253 for the training-average temperature profile. The joint gate obtains 61.54% conditional SC@8 versus 62.24% for fixed $T = 0.9$; the unrounded difference is -0.69 points, versus -0.15 for a same-switch-count random gate. These results delimit this architecture and training setup, not temperature opportunity or all possible predictors.

Temperature-pair heterogeneity. All 28 pairs were computed and retained. Table 21 shows the actual 0.5–0.9 switch, adjacent middle-temperature pairs, and several wider contrasts selected for interpretation rather than efficacy claims.

The largest listed correlations involve extreme temperatures, but this does not make them good actions: $T = 1.5$ has only about 45% aggregate accuracy in these labels. Reliably detecting a harmful contrast is different from identifying a question on which switching away from $T = 0.9$ improves accuracy.

Same-sample maximization. An outcome-assisted oracle that chooses the empirical best temperature and is evaluated on the same half appears to gain 20.53 points on fresh fit questions. Keeping the chosen actions fixed and evaluating on the other half reduces the two-direction average to 1.95 points. The corresponding calibration figures are 19.29 and 3.57 points and are direction-sensitive.

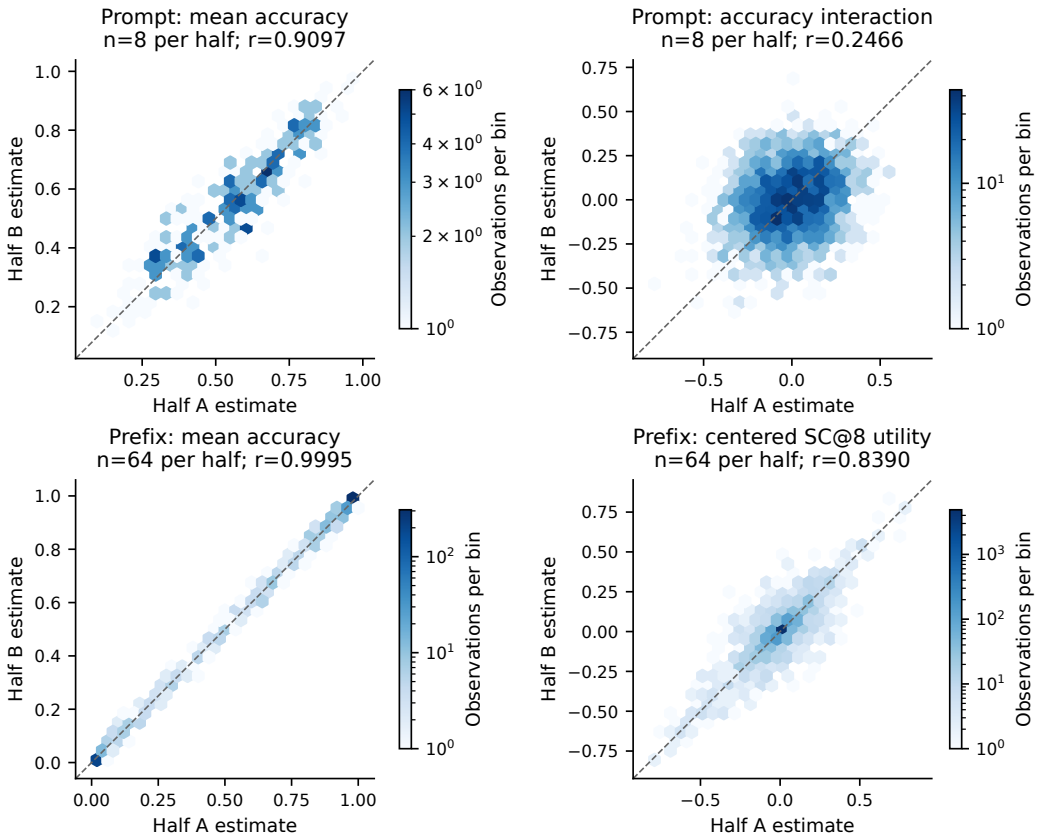


Figure 15: Split-half relationships from saved MATH responses. Hexagon color shows observation counts; dashed lines indicate equality, not fitted models. Top: 263 fresh-label prompt fit questions, eight responses per temperature per half. Bottom: 800 prefixes from 400 training questions, 64 responses per half. Mean-accuracy panels use one point per question or prefix; response panels use one point per state–temperature cell. Bottom right uses centered conditional SC@8, not the doubly centered accuracy at top right. Different states, budgets, and targets prevent a controlled prompt-versus-prefix interpretation. Correlations are descriptive point estimates.

Table 21: Split-half correlation of per-question accuracy differences for selected temperature pairs. Higher repeatability does not imply that either temperature is a useful action.

Temperature contrast	Fresh fit r	Calibration r
0.5 vs. 0.9	0.0623	0.0858
0.5 vs. 0.7	0.0312	-0.0142
0.7 vs. 0.9	-0.0298	-0.0446
0.1 vs. 0.9	0.3034	0.2241
0.1 vs. 1.3	0.4630	0.3881
0.1 vs. 1.5	0.5099	0.4334

This is a diagnostic of winner's-curse amplification, not the performance of the prompt-only MLP: the oracle observes 8 temperatures $\times$ 8 labeled responses per question.

G ADDITIONAL SELECTORS AND ACTION-SPACE ABLATIONS

These experiments document alternative selectors and action spaces. Unless explicitly stated, their policies, populations, and saved response pools differ from the final prompt-only MLP evaluated in Appendix E.

G.1 EARLIER PREFIX-BASED SELECTORS

A prefix-based random forest uses saved low-temperature prefixes and PCA6 hidden features on an outcome-screened cohort: individual accuracy increases from 54.62% to 57.60% (+2.99 points), while Vote@8 decreases from 73.14% to 67.99% (-5.14 points). These saved-continuation results motivate advantage supervision; they are neither prompt-only results nor evidence of a label-free deployment policy.

A further prefix-based comparison applies the no-PCA vote-margin MLP after cached $T = 0.1$ prefixes on 24 held-out screened MATH questions. It obtained 46.16% individual accuracy versus 49.50% for development-selected fixed $T = 0.9$. Retraining the same representation directly on centered correctness over all 84 available screened questions also remained below fixed temperature in grouped OOF evaluation (49.75% versus 51.84%). These comparisons use development data and saved response pools. They motivate the prompt-only decision setting but are not new untouched evaluations.

G.2 PREFIX-BASED RANDOM FOREST: MODAL AND FREQUENCY CONTROLS

The preceding hidden-state random forest selects a continuation temperature after a saved low-temperature prefix. Its most frequent action is $T_{\mathrm{mode}} = 0.1$ (27 of 45 prefixes), while its development-selected eight-action fixed comparator is $T^* = 0.9$. Table 22 shows positive gain over the modal policy under both endpoints, while the modal policy itself underperforms the fixed comparator, especially for conditional Vote@8.

Table 22: MATH prefix-based random-forest gain decomposition (percentage points). These comparisons use a different policy and population from the prompt-only MLP. Columns follow Equation 7.

Endpoint	Common	Beyond mode	Total vs. fixed 0.9
Individual accuracy	-2.81	+5.80	+2.99
Conditional Vote@8	-17.28	+12.14	-5.14

The same recorded analysis also preserves action frequencies within source and randomizes which prefix receives which temperature. Against this frequency- matched reference, contextual assignment contributes +4.91 individual-accuracy points (95% interval $[+0.31, +9.05]$) and +5.86 conditional-Vote@8 points ($[-0.25, +11.26]$). These two controls answer different questions: Table 22 compares with one modal temperature, whereas the frequency control preserves the entire action mixture.

G.3 A LARGER QUESTION-LEVEL FREQUENCY DECOMPOSITION

A separate MATH experiment using saved response pools evaluates a prompt-only random forest on 400 question-held-out predictions with one saved SC@8 pool per action. Both the question policy and validation-selected fixed $T = 0.9$ score 68.75%. Preserving the learned action frequencies within folds while randomizing their assignment to questions scores 67.29%, yielding the exact decomposition

$$0.00 \text{ pp} = \underbrace{-1.46 \text{ pp}}_{\text{action mixture}} + \underbrace{1.46 \text{ pp}}_{\text{question–temperature matching}} . \tag{8}$$

The matching interval is $[-0.48, +3.38]$ points. This larger cached diagnostic again suggests positive conditional matching offset by global calibration, but it uses a 15-action PCA random forest, earlier development questions, and a single eight-response pool; it is not an independent confirmation of the final MLP.

G.4 THREE-ACTION OUTPUT ABLATION

To test whether the eight-way output alone caused conservative selection, we restricted the original MLP to $\{0.1, 0.9, 1.3\}$ and separately fit one new three-output MLP with the same features, centered target, gate, roles, and seed. Both policies selected $T = 0.9$ on all 45 MATH evaluation questions, making them identical to fixed temperature and fixing their individual-accuracy difference at zero without new generation. On the common three-dimensional calibration target, the restricted original model has MSE 0.01589 and the new model 0.01620. Thus reducing the action head did not recover heterogeneous actions in this configuration; it does not rule out other action subsets or architectures.

H MATCHED CONTROLS FOR PREFIX-BASED SELECTION

This ablation uses a separate generation block from Table 3 and a selector trained on prefix features. On screened GSM8K questions, full-answer fixed $T = 1.1$ attains 62.92% individual accuracy; generating a 64-token $T = 0.1$ prefix and continuing at the same $T = 1.1$ gives 54.58%. Using the prefix selector gives 52.92%. Differences computed before rounding are -8.33 points for prefix generation and -1.67 for selection given the same prefixes. On OlympiadBench, the corresponding differences are +2.46 and -5.13 points. A low-temperature prefix is therefore not uniformly harmful. This is not a timing-only ablation of an otherwise identical prompt MLP.

For the same evaluation question set, define $A_0(T^*)$ as measured individual accuracy when the complete answer is generated at T^*. Define $A_{\mathrm{pre}}(T^*)$ as accuracy when a prefix is generated at $T = 0.1$ and its suffix at T^*, and $A_{\mathrm{pre}}(g)$ when a prefix-based selector g chooses the suffix temperature. The latter two policies use the same sampled prefixes. All three scores use the same question weighting; the two prefix scores also average over the sampled prefixes for each question. The measured total difference satisfies

$$A_{\mathrm{pre}}(g) - A_0(T^*) = \underbrace{A_{\mathrm{pre}}(T^*) - A_0(T^*)}_{\text{prefix-generation contrast}} + \underbrace{A_{\mathrm{pre}}(g) - A_{\mathrm{pre}}(T^*)}_{\text{selection given the same prefixes}} . \tag{9}$$

This identity specifies two controlled policy comparisons, rather than deriving the superiority of a selection architecture. The first contrast holds the suffix temperature fixed while changing the initial generation policy. The second holds the prefixes fixed while changing how suffix temperatures are chosen. Both are finite-sample accuracy estimates. The prefix selector g uses a different feature setting from the prompt-only MLP; this comparison therefore cannot isolate the effect of additional prefix information under an otherwise identical predictor.