RETHINKING ACTIVATION STEERING FOR MATHEMATICAL REASONING: WHAT ACTUALLY DRIVES ITS REASONING GAINS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering can improve mathematical reasoning and is commonly interpreted as controlling semantically meaningful internal reasoning states, such as reflection, reasoning depth, or over- and underthinking. We question whether such internal-state-control explanations are necessary to account for the resulting accuracy gains. Across three steering constructions, two model scales, and two mathematical reasoning benchmarks, we find that methods with different internal-state interpretations converge on a common downstream effect: they alter local decoder preferences and redirect generation among reasoning trajectories already supported by the model. Correct solutions and related reasoning strategies are frequently accessible under unsteered sampling. More directly, ordinary stochastic decoding whose parameters are selected to match steering's aggregate departure from greedy decoding achieves comparable accuracy while substantially reducing the remaining marginal benefit of steering; whole-thought exploration reproduces the same pattern. These results support a unified *greedy-relaxation* account of reasoning-oriented activation steering. Under this view, its accuracy gains need not be explained by a method's ability to identify and structurally control a particular internal reasoning state. The experiments instead localize a shared accuracy-relevant bottleneck at trajectory selection: steering changes which already-supported trajectory is realized, and ordinary stochastic decoding can exploit this bottleneck without reproducing the steering-specific activation intervention. This separates the internal representation targeted by a steering method from the mechanism explaining its performance gain.

1 INTRODUCTION

Activation steering has emerged as an effective way to control reasoning behavior at inference time. By adding a fixed or dynamically scaled direction to intermediate activations, steering methods can modulate reflection, verification, backtracking, reasoning depth, and computational effort, sometimes improving mathematical reasoning accuracy without retraining model weights (Rimsky et al., 2024; Chen et al., 2025; Li et al., 2026; Zhang et al., 2025). What remains less clear is *why* these interventions improve accuracy.

A prominent interpretation of reasoning-oriented activation steering is an *internal-state control* explanation. Recent methods identify and intervene on internal representations associated with particular reasoning modes. SEAL models transitions among execution, reflection, and transition states and steers their latent representations (Chen et al., 2025); ReBalance represents overthinking and underthinking as distinct hidden-state reasoning regimes and dynamically regulates between them (Li et al., 2026); and RISER constructs activation directions interpreted as latent cognitive primitives and adaptively composes them during reasoning (Ye et al., 2026). Despite their different implementations, these methods share a common explanatory pattern: steering regulates an internal reasoning state or mode, and the resulting behavioral change is used to explain improved task performance. Such control can produce recognizable output-level effects—for example, altered reflection, reasoning persistence, or corrective behavior—but these behaviors need not themselves be the direct causal source of improved reasoning accuracy.

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.

Recent work has also raised concerns about interpreting an identifiable activation pattern as a functionally privileged reasoning state: representations that appear strongly separable under probing can reflect task or format confounds and need not support the corresponding causal steering effect (Sahoo et al., 2026). More fundamentally, regardless of which internal representation a steering method manipulates, any improvement in final-answer accuracy must ultimately be realized through the sequence of continuations generated during inference. This makes trajectory selection a natural downstream interface shared across otherwise different steering mechanisms. Long-form reasoning is intrinsically multi-path: the same problem may admit multiple distinct reasoning trajectories, including different derivations that all lead to a correct solution. Greedy decoding, however, realizes only one such path by repeatedly selecting the locally highest-probability continuation. A single decoded sequence need not faithfully represent the probability mass distributed across a high-dimensional sequence distribution (Eikema & Aziz, 2020). Consistent with this mismatch, self-consistency improves reasoning by sampling multiple chains and aggregating their answers (Wang et al., 2023), and sampling can realize correct trajectories that greedy decoding misses without changing model parameters. This suggests that part of the observed reasoning deficit may arise from how learned capabilities are realized at inference time rather than from the absence of a correct solution in the model itself.

We therefore propose a competing *greedy-relaxation* hypothesis. Under this view, the central bottleneck lies not in the absence of an appropriate reasoning state, but in how the model's existing reasoning capability is realized at inference time. A reasoning model may support multiple viable solution trajectories, whereas greedy decoding commits generation to the single trajectory induced by successive local top-1 decisions. Activation steering can improve accuracy by altering some of these decisions and redirecting generation toward alternative trajectories already supported by the model, changing which existing reasoning possibility is realized. This creates a functional connection to sampling: sampling relaxes greedy commitment by allowing non-top-1 continuations to be selected, whereas activation steering does so by changing which continuation is selected under deterministic decoding. The two interventions need not induce the same trajectories or implement the same inference policy, but they may exploit the same trajectory-selection bottleneck—relaxing commitment to a single greedy trajectory.

This hypothesis differs from an internal-state control explanation in where it locates the accuracy-relevant bottleneck. Rather than requiring a particular latent reasoning mode, the greedy-relaxation hypothesis predicts that much of the steering benefit can be reproduced by changing how existing continuations are selected during generation. This yields two primary tests. First, if the relevant effect of steering is the extent to which it relaxes local greedy commitment, then ordinary stochastic decoding whose sampling parameters are selected to match steering's aggregate greedy-departure profile should achieve accuracy comparable to steering, without reproducing the particular direction-specific continuation selected by steering. Second, if steering and stochastic decoding exploit the same trajectory-selection opportunity, their gains should be strongly sub-additive: once matched stochastic decoding is already present, little marginal accuracy gain should remain for native steering.

We test these predictions on DeepSeek-R1-Distill-Qwen-1.5B and DeepSeek-R1-Distill-Qwen-7B, using MATH500 and GSM8K, and across three complementary steering constructions: SEAL, REBALANCE, and CPS. At the locations where each steering method would act, we use ordinary temperature-scaled top-p sampling and select its parameters to match the distribution of greedy-departure depths induced by that steering method. This tests whether a comparable relaxation of greedy commitment is sufficient even when stochastic decoding can realize different non-greedy alternatives across runs. We then test the same mechanism at a coarser scale with whole-thought exploration, where sampling is triggered by the unsteered model's local uncertainty.

Across models, datasets, and steering constructions, the results consistently support this trajectory-selection account. Correct solutions and closely related reasoning strategies are already accessible under unsteered sampling; steering-induced changes in textual decisions can carry much of the downstream accuracy benefit without requiring persistent propagation of the steered internal state; and stochastic decoding matched in its degree of greedy departure achieves accuracy comparable to steering while sharply reducing the remaining marginal benefit of steering. Whole-thought exploration reproduces the same pattern of accuracy matching and attenuated marginal steering gains at a coarser level of exploration. Together, these findings indicate that otherwise different steering methods can exploit a common trajectory-selection opportunity: they change which already-supported reasoning

trajectory is realized, rather than requiring a method-specific latent reasoning state to persist as the source of improved accuracy.

2 Related Work

Activation steering and representation engineering show that manipulating internal activations can control LLM behavior (Turner et al., 2024; Zou et al., 2023; Li et al., 2023; Rimsky et al., 2024; Todd et al., 2024; Liu et al., 2024). Reasoning-specific work extends this idea to execution/reflection states, over-/underthinking modes, latent skills, and reasoning effort or depth (Chen et al., 2025; Li et al., 2026; Zhang et al., 2025; 2026; Sheng et al., 2025; Liu et al., 2025; Ye et al., 2026). Despite differing constructions, these approaches commonly explain gains through regulation of internal variables representing the reasoning process. Related probing work likewise identifies structured latent features, while also showing that representational separability need not establish functional mechanism (Marks & Tegmark, 2024; Sahoo et al., 2026). We ask whether such internal-state control is necessary to explain the accuracy gain, or whether steering and stochastic decoding share a downstream bottleneck in trajectory selection.

3 Experimental Setup

Models and benchmarks. We evaluate DeepSeek-R1-Distill-Qwen-1.5B and DeepSeek-R1-Distill-Qwen-7B on the full MATH500 benchmark and a fixed 500-example GSM8K evaluation subset. Within each model–dataset setting, compared conditions use the same prompt, generation budget, answer extraction procedure, and correctness criterion. Unless an experiment requires an explicit parameter-selection split, evaluation is performed on the full evaluation set. Dataset-population and generation-budget details are given in Appendix A.1.

Steering constructions. We study three activation-steering constructions with distinct internal-state formulations: SEAL, which targets labeled execution, reflection, and transition states (Chen et al., 2025); ReBalance, which dynamically regulates confidence-dependent reasoning regimes (Li et al., 2026); and Cluster-Prototype Steering (CPS), which combines latent cluster prototypes without assigning them prespecified semantic labels. All method-specific steering components are constructed independently of the reported benchmark evaluation examples. SEAL and ReBalance use their official implementations; construction data separation and CPS details are given in Appendices D.1 and D.2.

Inference and evaluation. Greedy decoding is the default policy for unsteered and activation-steered conditions unless stated otherwise. Sampling experiments use the model's standard reasoning-time sampling configuration, except where decoding parameters are explicitly selected by an experiment. Randomized conditions are evaluated over multiple decoding seeds where applicable, and paired confidence intervals are computed over benchmark problems. Experiment-specific calibration, replay, and statistical protocols are described with the corresponding experiments and in the appendix.

4 Existing Correct Alternatives Beyond Greedy

We first ask whether the correct answers and reasoning patterns reached by steering are already accessible from the unsteered model when generation is allowed to move beyond a single greedy trajectory. We examine this at the levels of benchmark accuracy, answer-level support, and reasoning content.

4.1 Sampling and Steering Both Outperform Greedy Decoding

We compare unsteered greedy decoding, greedy decoding under each activation-steering method, and 16 independent full-evaluation-set runs of the model's standard reasoning-time sampling policy. Each sampling run produces one complete solution for every benchmark example; additional sampling-protocol details are provided in Appendix A.2.

Sampling improves over unsteered greedy decoding in all four model–dataset settings, with gains ranging from 0.56 to 13.00 percentage points, and achieves accuracy comparable to activation steering.

Table 1: **Benchmark accuracy under greedy decoding, sampling, and activation steering.** Sampling reports mean ± SD over 16 independent full-evaluation-set runs; steering conditions use greedy decoding.

Model	Dataset	Accuracy (%)				
		Greedy	Sampling (mean ± SD)	SEAL	ReBalance	CPS
1.5B	MATH500	67.6	80.60 ± 0.98	76.6	74.2	78.4
1.5B	GSM8K	71.0	77.54 ± 1.50	78.2	75.0	78.4
7B	MATH500	85.8	89.83 ± 0.92	88.8	88.4	89.2
7B	GSM8K	87.6	88.16 ± 0.79	88.2	88.2	88.4

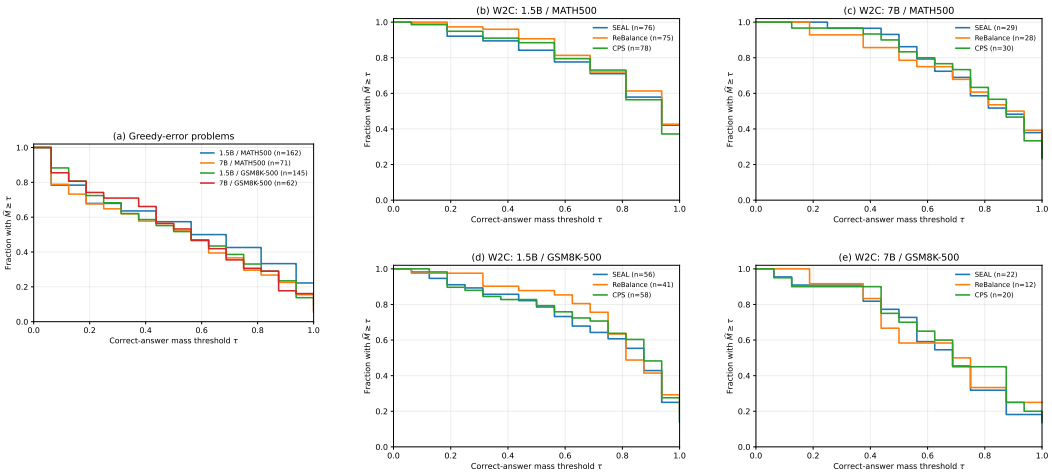

Figure 1: **Correct solutions retain substantial probability mass beyond greedy decoding.** Empirical survival curves of correct-answer mass $\widehat{M}$ for (a) greedy-error problems and (b–e) steering W2C examples across model–dataset settings.

Thus, substantial performance gains are already accessible by changing how the model's learned sequence distribution is realized at inference time, without changing model parameters or activations.

4.2 Correct Alternatives Remain Supported Beyond Greedy

We next focus on DeepSeek-R1-Distill-Qwen-1.5B on MATH500 and ask how much correct-answer probability remains available on problems missed by greedy decoding. Let π_{samp} denote the model's standard reasoning-time sampling policy. For a problem x_i, we estimate its correct-answer mass as

$$\widehat{M}_i = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}\left[Y_i^{(k)} = y_i^\star\right], \tag{1}$$

where $Y_i^{(k)}$ is the final answer produced by the k-th independent unsteered sample. We evaluate this quantity over greedy-error problems and over each steering method's wrong-to-correct (W2C) subset, consisting of problems missed by greedy decoding but corrected by steering. For all greedy-error problems, $K = 16$, using the 16 full-benchmark unsteered sampling runs above. For W2C cases, we generate 16 additional unsteered rollouts per problem, giving $K = 32$ in total. The population definitions and survival-curve construction are detailed in Appendix A.4.

Figure 1 shows the resulting empirical survival curves for the 1.5B model on MATH500. Greedy errors retain substantial correct-answer probability under finite unsteered sampling, and this support is especially strong among the cases rescued by steering. Thus, steering frequently corrects problems for which a correct solution is already present in the model's unsteered sampling distribution.

Answer-level support does not establish that steering reaches a correct answer through reasoning that is also present among unsteered samples. We therefore test this directly at the thought and full-solution levels.

4.3 Steering Reasoning Has Counterparts Under Finite Sampling

Thought-level counterparts. We analyze all 78 CPS W2C cases on MATH500. At the first reasoning boundary where CPS changes the unsteered greedy choice, we generate the CPS thought and 16 unsteered sampled alternatives from the identical prefix and cache. The reasoning-boundary convention is specified in Appendix A.3. To discount similarity inherited from the shared prefix, we compare hidden-state displacements from the common event-start state and calibrate the nearest CPS-to-sample similarity against the natural nearest-neighbor similarities among the 16 unsteered thoughts themselves. The nearest sampled thought has mean cosine similarity 0.892, and $69/78$ cases (88.5%) fall within the case-specific natural sampled neighborhood. Only $6/78$ cases (7.7%) contain an exact textual copy, whereas $74/78$ (94.9%) contain at least one sampled thought judged to follow an equivalent mathematical plan. Thus, the correspondence is primarily semantic and representational rather than literal reproduction. The representation, natural-neighborhood, and semantic-counterpart definitions are given in Appendix A.5.

Solution-level counterparts. We next compare complete mathematical solutions. For 76 of the 78 W2C cases, a fixed pool of 32 unsteered rollouts contains at least eight distinct correct, closed, and non-truncated solutions; we pair the CPS solution with the first eight eligible solutions. Two cases are excluded because fewer than eight eligible solutions are available. Across the resulting 608 blinded pairs, 566 (93.1%) are judged to use the same core mathematical strategy. At the case level, $75/76$ CPS solutions (98.7%) have at least one strategy counterpart among the eight eligible unsteered solutions. The sole exception is an unclosed CPS trajectory that reaches the 10,000-token generation limit. The full selection and judging protocol is provided in Appendix A.6.

Together, these results show that the correct answers and reasoning strategies reached by steering are frequently already accessible under finite sampling from the unsteered model. Steering therefore need not create new mathematical capability to explain its accuracy gains; it can instead change which already-supported trajectory is realized. The next section asks how an activation-space intervention produces this change.

5 Activation Steering Acts Through Decoder-Level Changes

Activation steering originates inside the network, but its durable behavioral effect must ultimately be expressed through changes in the decoder distribution, selected tokens, and the textual prefixes that determine subsequent generation. We therefore ask whether steering systematically changes local greedy decisions, whether these decoder-level effects are more functionally informative than raw activation-space geometry, and whether the long-horizon accuracy gain still requires the steered hidden state or KV cache after the altered text has been committed. We address these questions with local decision analysis across steering methods, a controlled CPS functional-equivalence case study, and clean-cache replay. Unless otherwise stated, these mechanistic analyses use a fixed 200-problem subset of MATH500, with paired conditions evaluated on the same problems.

5.1 Steering Perturbs Local Greedy Decisions

At fixed generation contexts with identical textual prefixes and KV caches, steering changes the greedy next-token choice at 9.18%–39.87% of reasoning boundaries across the six model–method configurations. These changes concentrate near weakly determined decisions: the unsteered top-1/top-2 margin predicts steering-induced argmax flips with AUROC 0.793–0.838. Steering also produces non-trivial redistribution of decoder scores, with mean symmetric KL ranging from 0.105 to 0.622. Full configuration-level diagnostics are reported in Appendix B.1, Table 4.

Table 2: **Clean-cache replay preserves the steering benefit.** Accuracy on the fixed 200-problem MATH500 subset. Token/segment replay retains the steering-selected text while discarding the steered cache; segment replay reconstructs a clean unsteered cache by teacher forcing.

Model	Method	Accuracy (%)			
		Unsteered greedy	Full steering	Token replay	Segment replay
1.5B	SEAL	31.67	61.67	58.33	60.83
1.5B	ReBalance	31.67	41.67	39.17	40.83
1.5B	CPS	46.50	68.50	66.50	68.50
7B	SEAL	80.00	80.83	83.33	84.17
7B	ReBalance	79.17	80.00	80.83	81.67
7B	CPS	80.00	82.50	81.67	82.50

5.2 Decoder Effects Better Capture Functional Equivalence

Raw activation-space similarity need not determine functional similarity because the same perturbation is transformed by the model in a context-dependent manner. In a controlled CPS case study, we characterize the local decoder effect of a steering vector v through its induced logit change,

$$\Delta z_v(s) = z_v(s) - z_0(s) \approx J_s v, \tag{2}$$

where J_s denotes the context-dependent Jacobian from the intervention layer to the decoder logits.

We construct alternative vectors that match the decoder effect of the original CPS direction while allowing their activation-space geometry to differ. On held-out contexts, a functionally matched vector has only 0.370 cosine similarity with the original direction, yet reaches 0.985 cosine similarity in induced logit changes and 94.03% agreement in greedy next-token decisions. Functionally matched vectors also retain most of the CPS accuracy gain on MATH500. Construction details, the relative logit-delta error, and full accuracy results are reported in Appendix B.2. These results indicate that the context-dependent decoder effect is a more informative functional description of the intervention than raw activation-vector geometry.

5.3 Clean-Cache Replay Localizes Long-Horizon Gains to Generated Text

Full activation steering changes both the current decoder distribution and the hidden states and KV cache propagated into future generation. We separate these channels with clean-cache action cloning. In *token replay*, an auxiliary steered forward pass determines only the next greedy token; the steered hidden state and cache are discarded before continuation. In *segment replay*, auxiliary steered generation produces one complete reasoning segment; its cache is discarded and the generated text is teacher-forced through the unmodified model to reconstruct a clean cache. Operational details are provided in Appendix B.3.

Across all three methods on the 1.5B subset, token and segment replay preserve most of the full-steering improvement, with segment replay closely matching or matching full steering. At 7B, where full-steering gains are smaller, both replay conditions remain comparable to full steering and often match or exceed it. A complementary CPS logits/cache factorial yields the same asymmetry (Appendix B.3, Table 5): preserving the steering-induced decoder effect with a clean cache retains most of the gain, whereas carrying the steered cache without the corresponding decoder effect does not.

Taken together, the evidence supports a compact causal picture:

$$\text{activation perturbation} \longrightarrow \text{decoder distribution and local choice} \longrightarrow \text{textual trajectory}.$$

Steering changes weakly determined local decisions; geometrically different perturbations can remain behaviorally similar when they induce similar decoder effects; and once an altered token or thought has been committed, most of the downstream accuracy benefit can be carried by the resulting text without continuously propagating a steering-specific hidden state or KV cache. The activation perturbation remains the cause of the local change; the clean-cache result localizes its durable downstream effect to the generated text and trajectory.

This motivates the next question: if the accuracy-relevant effect is expressed through trajectory selection, can the comparable accuracy gains be obtained by changing the decoding policy without activation steering?

6 Stochastic Decoding Tests the Greedy-Relaxation Mechanism

The preceding section shows that steering's durable behavioral effect can be carried by changes in decoder actions and the textual trajectories they induce. The greedy-relaxation hypothesis therefore makes a stronger prediction: if the accuracy gain mainly comes from relaxing local top-1 commitment, ordinary stochastic decoding should obtain comparable gains when it produces departures from greedy of comparable strength, without repeatedly realizing the same direction-specific continuation. We test this prediction with greedy-departure matching and then with whole-thought exploration.

6.1 Greedy-Departure Matching Tests Trajectory-Selection Sufficiency

At the locations where each steering method would act, we use ordinary temperature-scaled top-p sampling and select its parameters so that the resulting departures from the unsteered greedy action have similar overall depth to those induced by steering; all other positions remain greedy. For an unsteered state s, let $a_G(s) = \arg\max_a z_0(a \mid s)$. We define the greedy-departure depth of any selected action a as

$$\delta(s, a) = z_0(a_G(s) \mid s) - z_0(a \mid s). \tag{3}$$

Steering induces a distribution $P_V(\delta)$ through its selected actions, whereas temperature/top-p sampling at the same locations induces $P_{T,p}(\delta)$. Because the top-1 action remains available under ordinary sampling, these distributions jointly capture how often generation leaves greedy and how far it moves when it does.

For each dataset, 100 examples are reserved for parameter selection and the remaining 400 are held out for evaluation. For each model, dataset, and steering method, we select

$$(T^\star, p^\star) = \arg\min_{T,p} W_1(P_V(\delta), P_{T,p}(\delta)), \tag{4}$$

on fixed unsteered prefixes at the intervention locations in the parameter-selection split, with problems weighted equally. Parameter-selection details, selected sampling parameters, and frozen-prefix matching diagnostics are reported in Appendix C.2; Appendix C.3 separately checks the realized departure profile after stochastic choices have changed the generated trajectory.

We compare four conditions: G (unsteered greedy decoding), V (native activation steering), S (greedy-departure-matched stochastic decoding), and $S + V$ (the same matched sampling policy combined with native steering). In $S + V$, the selected (T, p) is applied at the same locations while the native activation intervention and its cache propagation remain active.

Figure 2 tests whether matching the degree of greedy departure is sufficient to achieve accuracy comparable to steering. Across the 12 evaluated configurations, the equal-weight difference $A(S) - A(V)$ is $+0.28$ percentage points, and matched stochastic decoding lies within 2 percentage points of native steering in 9 of 12 configurations. Because sampling can realize different non-greedy continuations across runs, this accuracy match does not require repeatedly selecting the same direction-specific continuation. Full configuration-level accuracies and paired effects are reported in Appendix C.1, Tables 6 and 7.

The second prediction concerns overlap in the source of the gain. Let $\Delta_V = A(V) - A(G)$ denote the native steering gain and $\Delta_{V|S} = A(S + V) - A(S)$ the residual steering gain after matched stochastic decoding. We summarize their difference by

$$I_S = \Delta_{V|S} - \Delta_V. \tag{5}$$

Across the 12 held-out matched-decoding configurations, the equal-weight native steering gain is $\Delta_V = +4.29$ percentage points, whereas the residual gain after matched stochastic decoding is only $\Delta_{V|S} = +0.01$ points. The corresponding interaction is $I_S = -4.28$ points and is negative in all 12 configurations. Thus, matched stochastic decoding achieves accuracy comparable to native steering and leaves substantially less marginal accuracy gain for steering. Together, these results support relaxation of greedy trajectory commitment, rather than a particular direction-specific internal-state transition, as the shared accuracy-relevant opportunity.

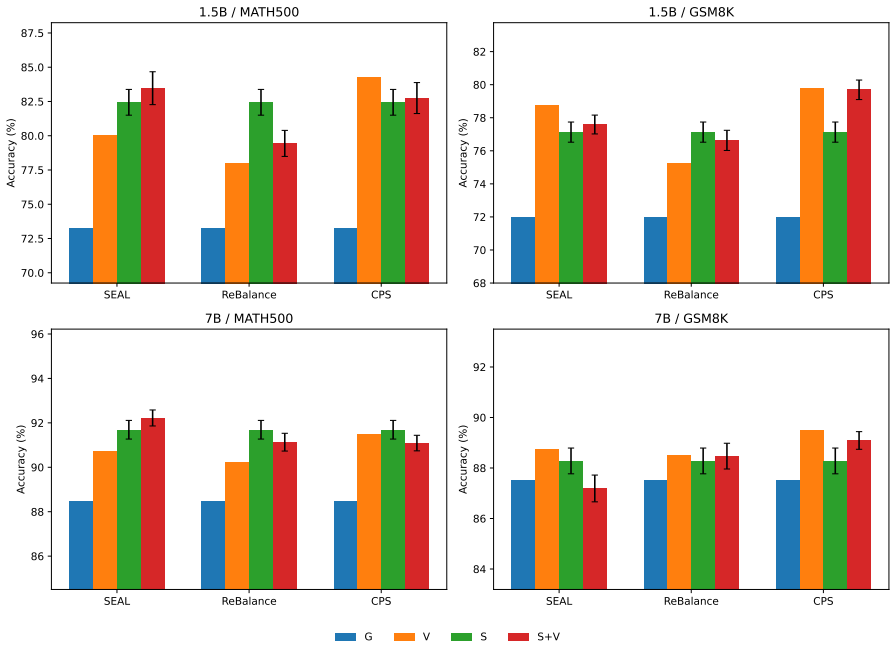

Figure 2: **Matched stochastic decoding achieves accuracy comparable to native steering.** Bars compare unsteered greedy (G), native steering (V), matched stochastic decoding (S), and $S + V$; stochastic error bars denote ± 1 SD over eight seeds.

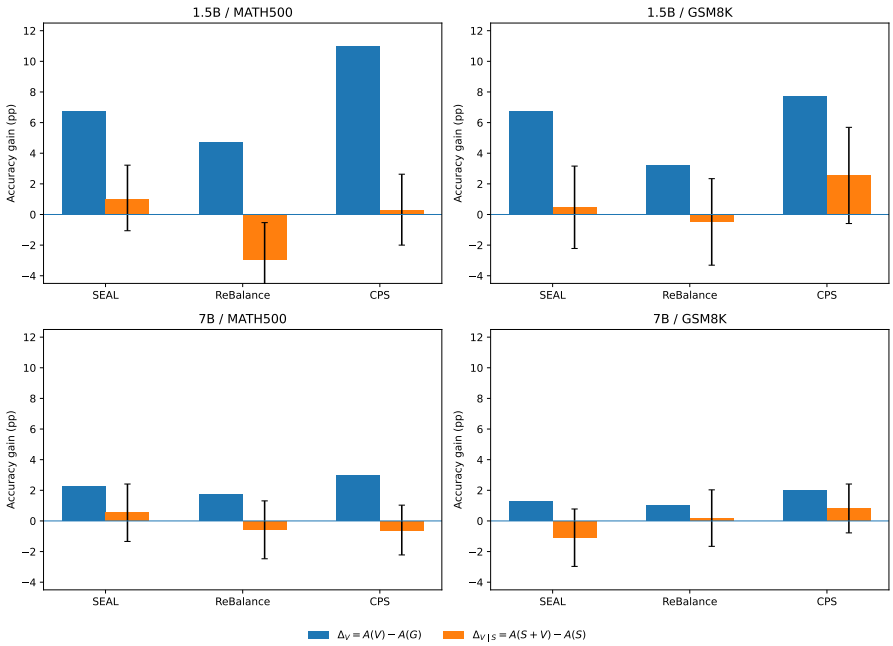

Figure 3: **Matched stochastic decoding reduces the marginal accuracy benefit of native steering.** Bars show $\Delta_V = A(V) - A(G)$ and $\Delta_{V|S} = A(S + V) - A(S)$; error bars on the latter denote 95% paired-bootstrap confidence intervals.

Table 3: **Aggregate effects of whole-thought exploration.** Values are percentage points; brackets give 95% paired-bootstrap CIs for equal-weight macro-averages. The native steering effect in this evaluation is $+4.52$ points (95% CI $[+3.19, +5.63]$).

Dose	$E - G$	$E - V$	$(E + V) - E$	**Interaction** I_{E_k}
q10	$+2.02$ $[+1.01, +3.39]$	-2.50 $[-3.37, -1.07]$	$+2.79$ $[+1.64, +3.93]$	-1.72 $[-3.07, -0.20]$
q25	$+4.63$ $[+3.58, +6.04]$	$+0.10$ $[-0.72, +1.51]$	$+0.41$ $[-0.77, +1.28]$	-4.11 $[-5.66, -2.68]$
q50	$+5.48$ $[+4.32, +7.02]$	$+0.96$ $[+0.11, +2.38]$	$+0.37$ $[-0.64, +1.46]$	-4.15 $[-5.60, -2.37]$
ALL	$+5.44$ $[+4.38, +7.11]$	$+0.92$ $[+0.22, +2.43]$	$+0.50$ $[-0.40, +1.63]$	-4.02 $[-5.39, -2.22]$

6.2 Whole-Thought Exploration Replicates Accuracy Matching and Gain Attenuation

We next ask whether the same pattern of accuracy matching and reduced marginal steering gain persists under a coarser uncertainty-triggered sampling policy. At each thought boundary, we compute the unsteered model's local top-1/top-2 margin,

$$m(s_t) = z_{(1)}(s_t) - z_{(2)}(s_t) = \log \frac{p_{(1)}(s_t)}{p_{(2)}(s_t)}. \tag{6}$$

We trigger whole-thought sampling when this margin falls below the 10th, 25th, or 50th percentile estimated on an independent parameter-selection split; under ALL, every subsequent reasoning segment is sampled. Otherwise the next reasoning segment is generated greedily. Exact model–dataset thresholds and the full policy are reported in Appendix C.4.1 and Table 10.

Let E_k denote whole-thought exploration at dose $k \in \{\text{q10}, \text{q25}, \text{q50}, \text{ALL}\}$, and let $E_k + V$ denote the same exploration policy combined with native steering. The exploration trigger is defined from the unsteered margin in both conditions. All contrasts are computed within each model–dataset–method configuration before taking an equal-weight macro-average; confidence intervals and bootstrap details are given in Appendix C.4.2. We define the interaction analogously as $I_{E_k} = \Delta_{V|E_k} - \Delta_V$.

Table 3 shows a clear dose-response. At q10, exploration recovers only part of the native-steering gain and leaves a substantial residual benefit for steering. By q25, exploration matches native steering on average, while the residual steering gain contracts sharply; at q50 and ALL, exploration modestly exceeds native steering on average and adding steering contributes little additional accuracy. The corresponding interactions are strongly negative from q25 onward. Configuration-level residual gains and interactions are reported in Appendix C.4.3, with the compact residual summary in Figure 5 and the full configuration panels in Figures 6 and 7.

The primary test remains greedy-departure matching: once ordinary stochastic decoding reproduces the aggregate extent to which steering leaves greedy, most of the accuracy gain is reproduced and native steering becomes strongly sub-additive. Whole-thought exploration shows that the same pattern of accuracy matching and gain attenuation survives under a substantially coarser uncertainty-triggered sampling policy, reinforcing a shared trajectory-selection bottleneck rather than an accuracy benefit that requires a steering-specific internal state to persist.

7 Conclusion

Across steering constructions, model scales, and mathematical benchmarks, our results support a trajectory-selection account of the accuracy gain. Correct solutions and related reasoning strategies are often already accessible under unsteered sampling; steering changes local decoder decisions, and most of its long-horizon benefit survives after the steered hidden state and cache are discarded once the resulting text is committed. Ordinary stochastic decoding can recover and compete for the same accuracy opportunity. Together, these findings suggest viewing reasoning-oriented activation steering as a structured, context-dependent inference policy for redirecting among already-supported trajectories, rather than requiring a method-specific latent reasoning state to persist as the source of improved accuracy.

Ethics Statement

This work studies inference-time steering and sampling behavior in language models on mathematical reasoning benchmarks. It does not involve human participants, private or sensitive personal data, or deployment decisions. The experiments use benchmark problems and model-generated outputs, and our analysis is limited to mechanistic questions about mathematical reasoning. Although activation steering and related inference-time control techniques could in principle be applied to alter model behavior in broader settings, we do not evaluate or advocate their use in safety-critical applications. We report both beneficial and harmful behavioral effects of the interventions, including correct-to-wrong transitions and changes in generation closure and length, to avoid presenting steering as uniformly beneficial or reliable. To the best of our knowledge, the study raises no additional human-subject, privacy, or sensitive-data concerns, and we adhere to the ICLR Code of Ethics.

Reproducibility Statement

We provide the information needed to reproduce the reported analyses throughout the main text and appendices. Section 3 specifies the models, benchmarks, steering constructions, and shared inference conventions. Appendix A documents the evaluation population, sampling protocol, reasoning-boundary convention, correct-answer-mass estimation, and thought- and solution-level counterpart analyses. Appendix B gives the local decoder diagnostics, functional-equivalence construction, and clean-cache replay protocol. Appendix C reports the configuration-level results, parameter-selection procedure, frozen- and on-policy greedy-departure diagnostics, whole-thought exploration thresholds, and bootstrap aggregation procedures. Finally, Appendix D.1 describes steering construction and data separation, with CPS-specific construction details in Appendix D.2. Tables report the random-seed counts and confidence-interval procedures used by the corresponding experiments.

AI use statement

In this work, we used generative AI tools, including ChatGPT, to assist with manuscript drafting and language editing, organization and restructuring of the paper, LaTeX formatting, literature and reference checks, preparation and revision of a scientific figure, and feedback on the presentation and interpretation of existing experimental analyses. The numerical results reported in the paper were taken from the authors' experimental outputs rather than generated by the AI tools. All AI-assisted text, citations, figures, and analytical suggestions were reviewed and revised by the authors against the underlying experiments and source materials. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Runjin Chen, Zhenyu Zhang, Junyuan Hong, Souvik Kundu, and Zhangyang Wang. SEAL: Steerable reasoning calibration of large language models for free. *arXiv preprint arXiv:2504.07986*, 2025. URL https://arxiv.org/abs/2504.07986.

Bryan Eikema and Wilker Aziz. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4506–4520, 2020. doi: 10.18653/v1/2020.coling-main.398. URL https://aclanthology.org/2020.coling-main.398/.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1797.

Yulin Li, Tengyao Tu, Li Ding, Junjie Wang, Huiling Zhen, Yixin Chen, Yong Li, and Zhuotao Tian. Efficient reasoning with balanced thinking. *arXiv preprint arXiv:2603.12372*, 2026. URL https://arxiv.org/abs/2603.12372.

Sheng Liu, Haotian Ye, Lei Xing, and James Y. Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 32287–32307. PMLR, 2024. URL https://proceedings.mlr.press/v235/liu24bx.html.

Sheng Liu, Tianlang Chen, Pan Lu, Haotian Ye, Yizheng Chen, Lei Xing, and James Zou. Fractional reasoning via latent steering vectors improves inference time compute. *arXiv preprint arXiv:2506.15882*, 2025. URL https://arxiv.org/abs/2506.15882.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

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

Subramanyam Sahoo, Vinija Jain, Aman Chadha, and Divya Chaudhary. Linear probes detect task format, not reasoning mode in language model hidden states. *arXiv preprint arXiv:2606.02907*, 2026. URL https://arxiv.org/abs/2606.02907.

Leheng Sheng, An Zhang, Zijian Wu, Weixiang Zhao, Changshuo Shen, Yi Zhang, Xiang Wang, and Tat-Seng Chua. On reasoning strength planning in large reasoning models. *arXiv preprint arXiv:2506.08390*, 2025. URL https://arxiv.org/abs/2506.08390.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2024. URL https://arxiv.org/abs/2308.10248.

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

Wencheng Ye, Xiaoyang Yuan, Yi Bin, Hengyu Jin, Liang Peng, Pengpeng Zeng, and Heng Tao Shen. RISER: Orchestrating latent reasoning skills for adaptive activation steering. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4627–4644, 2026. doi: 10.18653/v1/2026.findings-acl.226. URL https://aclanthology.org/2026.findings-acl.226/.

Dongxu Zhang, Yiding Sun, Zihao Guo, Xiangyang Yang, Kai Tang, Lin Chen, Cheng Tan, and Jihua Zhu. SPARK: Susceptibility-guided profiling and steering of latent reasoning states in large language models. *arXiv preprint arXiv:2607.10296*, 2026. URL https://arxiv.org/abs/2607.10296.

Zhenyu Zhang, Xiaoxia Wu, Zhongzhu Zhou, Qingyang Wu, Yineng Zhang, Pragaash Ponnusamy, Harikaran Subbaraj, Jue Wang, Shuaiwen Leon Song, and Ben Athiwaratkun. Understanding and steering the cognitive behaviors of reasoning models at test-time. *arXiv preprint arXiv:2512.24574*, 2025. URL https://arxiv.org/abs/2512.24574.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

A Additional Analyses for Existing Correct Alternatives

A.1 Evaluation Population and Generation Conventions

MATH500 evaluation uses the complete standard 500-problem benchmark. For GSM8K, we use a fixed 500-example evaluation subset selected to approximately balance problem difficulty and problem type; the same subset is reused across experiments. Stochastic experiments use experiment-specific random seeds as specified by their respective protocols. All main experiments allow up to 10,000 newly generated tokens per problem. Method-specific decoding and stopping behavior follows the corresponding implementation or experiment protocol.

A.2 Benchmark Sampling Protocol

For the benchmark comparison in Section 4.1, each model–dataset setting is evaluated with 16 independent runs of the model's official reasoning-time sampling configuration. These sampling parameters are not selected using benchmark accuracy and are distinct from the temperature/top-p parameters selected later for greedy-departure matching. Each run generates one complete solution for every example in the evaluation set, yielding one evaluation-set accuracy. The sampling entry in Table 1 is the mean and standard deviation of these 16 accuracies. The unsteered and activation-steering conditions use greedy decoding under their corresponding inference procedures.

A.3 Reasoning-Boundary Convention

For boundary-based analyses and CPS interventions, a *reasoning boundary* denotes the tokenizer token representing the consecutive-newline thought-split separator while generation remains inside the `<think>...</think>` region. The boundary rule does not trigger after the model leaves the thinking region and enters the answer stage. This convention is used when locating CPS intervention sites and when extracting thought-level counterparts.

A.4 Correct-Answer Mass Estimation

The correct-answer-mass analysis in Section 4.2 is conducted on DeepSeek-R1-Distill-Qwen-1.5B on MATH500. For each problem x_i, independent unsteered rollouts are evaluated with the same answer-extraction and correctness criterion used for benchmark accuracy. The empirical correct-answer mass is

$$\widehat{M}_i = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}\left[Y_i^{(k)} = y_i^\star\right]. \tag{7}$$

For the population of all greedy-error problems, $K = 16$, corresponding to the 16 evaluation-set sampling runs used in Section 4.1. For each method-specific W2C case, we generate 16 additional unsteered rollouts, so the mass estimate uses $K = 32$ in total. A method-specific W2C set contains problems answered incorrectly by its paired unsteered greedy condition and correctly by that steering method. For any analyzed population $\mathcal{P}$, the survival curve plotted in Figure 1 reports

$$\widehat{S}(\tau) = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \mathbb{1}\left[\widehat{M}_i \geq \tau\right]. \tag{8}$$

Thus, the curve measures how frequently a specified amount of correct-answer probability remains accessible under finite unsteered sampling.

A.5 Thought-Level Counterpart Analysis

The thought-level analysis in Section 4.3 uses all 78 CPS W2C cases on MATH500. For each case, we locate the first reasoning boundary at which CPS changes the unsteered greedy next-token choice, using the convention in Appendix A.3. From the identical textual prefix and KV cache at that boundary, we obtain the next CPS-generated thought and 16 thoughts sampled from the unsteered model.

Table 4: **Activation steering directly perturbs local greedy decisions.** Results are computed on the fixed 200-problem MATH500 mechanism subset. Argmax flips are measured from paired unsteered and steered forward passes at the same reasoning state. Low-margin AUROC measures how well the unsteered top-1/top-2 margin predicts a steering-induced argmax change.

Model	Method	Boundary sites	Argmax flip (%)	Low-margin AUROC	Mean sym. KL
1.5B	SEAL	6,789	27.37	0.814	0.374
1.5B	ReBalance	6,679	11.12	0.818	0.105
1.5B	CPS	10,230	39.87	0.793	0.622
7B	SEAL	5,939	25.49	0.838	0.291
7B	ReBalance	5,939	9.18	0.805	0.169
7B	CPS	8,630	34.62	0.812	0.487

To reduce similarity inherited from the shared prefix, each thought is represented by its hidden-state displacement from the common event-start representation:

$$\Delta h_i^V = h_{i,\mathrm{end}}^V - h_{i,\mathrm{start}}, \qquad \Delta h_{ij}^S = h_{ij,\mathrm{end}}^S - h_{i,\mathrm{start}}, \tag{9}$$

where V denotes the CPS thought and S an unsteered sampled thought. The closest sampled representation is summarized by

$$c_i^{\max} = \max_{1 \le j \le 16} \cos\big(\Delta h_i^V, \Delta h_{ij}^S\big)\,. \tag{10}$$

We calibrate this similarity against the natural variation among the 16 unsteered thoughts from the same prefix. For each sampled thought, we compute its leave-one-out nearest-neighbor cosine among the other 15 samples. The case-specific natural-neighborhood threshold T_i is the 10th percentile of these 16 nearest-neighbor similarities. A CPS thought has a representation-level counterpart when $c_i^{\max} \ge T_i$. Exact-text counterparts are identified by literal equality of the generated thought text, while semantic counterparts are determined by whether at least one sampled thought follows an equivalent mathematical plan.

A.6 Solution-Level Counterpart Analysis

For the solution-level counterpart analysis in Section 4.3, we begin from a fixed pool of 32 unsteered rollouts for each of the 78 CPS W2C cases on MATH500. An unsteered rollout is eligible only if it is correct, closed, non-truncated, and textually distinct from the other selected solutions. For 76 cases, at least eight such solutions are available; we pair the CPS solution with the first eight eligible solutions in the fixed pool. Cases 43 and 491 are excluded because they contain only five and one eligible solutions, respectively, yielding $76 \times 8 = 608$ judged pairs.

The judge is blinded to which solution came from steering, and pair order is randomized. A pair is labeled as sharing a plan when the two solutions use the same core mathematical strategy; differences in wording, verbosity, local algebraic manipulation, or minor execution details are not sufficient to make the plans different. A CPS solution is counted as having a sampled plan counterpart when at least one of its eight eligible unsteered solutions receives this label.

B Additional Analyses for Decoder-Level Changes

B.1 Local Decision Diagnostics

The local-decision analysis in Section 5.1 uses the fixed 200-problem MATH500 mechanism subset. At each reasoning boundary, unsteered and steered forward passes are evaluated from the same textual prefix and KV cache. We record whether steering changes the greedy next-token argmax, the unsteered top-1/top-2 logit margin, and the symmetric KL divergence between the paired next-token distributions. Low-margin AUROC measures how well the unsteered top-1/top-2 margin predicts a steering-induced argmax change.

Table 5: **CPS logits/cache factorial.** Accuracy is measured on MATH500. The comparison separates the decoder effect used for the current choice from the cache propagated into subsequent generation.

Condition	Logits	Cache	Accuracy (%)
Unsteered	Unsteered	Unsteered	73.25
Steering logits	Steering	Unsteered	82.00
Steering cache	Unsteered	Steering	67.25
Full steering	Steering	Steering	84.25

B.2 Functional-Equivalence Analysis

The CPS functional-equivalence experiment in Section 5.2 compares activation-space geometry with the decoder effect induced on a finite set of generation contexts. For a steering vector v, the relevant local quantity is the induced logit change $\Delta z_v(s) = z_v(s) - z_0(s)$, approximated by the context-dependent map $J_s v$ in Eq. 2. Alternative vectors are constructed to match this decoder effect while allowing their activation-space directions to differ.

On held-out contexts, a functionally matched vector has cosine similarity 0.370 with the original CPS direction but cosine similarity 0.985 between their induced logit changes. Its relative logit-delta error is 0.174, and the two interventions agree on the greedy next-token choice at 94.03% of evaluated contexts. The correspondence also extends to full-generation accuracy on MATH500: the original CPS intervention reaches 84.25%, the closest functionally matched vector reaches 83.75%, and other vectors constructed under the same decoder-effect matching objective reach 79.75%–82.25%, compared with a 73.25% unsteered baseline. Because the matching constraint is imposed on a finite set of states, these results establish local functional equivalence rather than global equivalence after generated trajectories diverge.

B.3 Clean-Cache Replay and Logits/Cache Factorial

The clean-cache experiments in Section 5.3 separate the effect of the generated text from persistent propagation of steered internal representations. For token replay, unsteered and steered forward passes are evaluated from the same current prefix; the steered pass determines the greedy next token, after which its hidden state and KV cache are discarded and continuation proceeds through the clean unmodified path. For segment replay, auxiliary steered generation produces a complete reasoning segment to the next boundary. The auxiliary steered cache is then discarded, and the generated segment is teacher-forced through the unmodified model to reconstruct a clean unsteered KV cache before continuation. The procedure is repeated at subsequent steering locations.

A complementary CPS factorial independently varies the decoder logits used for the current choice and the cache propagated forward. The resulting accuracies are shown in Table 5.

The factorial reproduces the asymmetry seen in clean-cache cloning: using the steering-induced decoder effect with an unsteered cache retains most of the full-steering gain, whereas retaining the steered cache without the corresponding decoder effect does not.

C Additional Analyses for Greedy-Relaxation Tests

C.1 Configuration-Level Greedy-Departure Results

Experiments use DeepSeek-R1-Distill-Qwen-1.5B and DeepSeek-R1-Distill-Qwen-7B on MATH500 and GSM8K with SEAL, ReBalance, and CPS. For each dataset, 100 examples are reserved for matching-parameter selection and 400 disjoint examples are used for held-out evaluation.

C.1.1 Evaluation Protocol

We evaluate four conditions: G (unsteered greedy decoding), V (native activation steering), S (greedy-departure-matched stochastic decoding), and $S + V$ (matched stochastic decoding combined with native steering). The stochastic conditions S and $S + V$ are each run with eight independent decoding seeds for every model–dataset–method configuration. G and V are deterministic reference

Table 6: **Configuration-level accuracy for greedy-departure-matched decoding.** G is unsteered greedy decoding, V native activation steering, S greedy-departure-matched stochastic decoding, and $S + V$ their combination. S and $S + V$ report mean accuracy $\pm$ standard deviation across eight independent decoding seeds; G and V are deterministic references.

Model	Dataset	Method	Accuracy (%)			
			G	V	S (mean $\pm$ SD)	$S + V$ (mean $\pm$ SD)
1.5B	MATH500	SEAL	73.25	80.00	82.44 ± 0.94	83.47 ± 1.20
1.5B	MATH500	ReBalance	73.25	78.00	82.44 ± 0.94	79.44 ± 0.95
1.5B	MATH500	CPS	73.25	84.25	82.44 ± 0.94	82.75 ± 1.13
1.5B	GSM8K	SEAL	72.00	78.75	77.13 ± 0.61	77.59 ± 0.57
1.5B	GSM8K	ReBalance	72.00	75.25	77.13 ± 0.61	76.63 ± 0.61
1.5B	GSM8K	CPS	72.00	79.75	77.13 ± 0.61	79.69 ± 0.59
7B	MATH500	SEAL	88.50	90.75	91.69 ± 0.42	92.22 ± 0.36
7B	MATH500	ReBalance	88.50	90.25	91.69 ± 0.42	91.13 ± 0.40
7B	MATH500	CPS	88.50	91.50	91.69 ± 0.42	91.09 ± 0.35
7B	GSM8K	SEAL	87.50	88.75	88.28 ± 0.51	87.19 ± 0.53
7B	GSM8K	ReBalance	87.50	88.50	88.28 ± 0.51	88.47 ± 0.51
7B	GSM8K	CPS	87.50	89.50	88.28 ± 0.51	89.09 ± 0.35

Table 7: **Configuration-level sufficiency and competition effects for greedy-departure-matched decoding.** $V - G$ is the native steering gain under greedy decoding; $S - V$ directly compares matched stochastic decoding with native steering; $(S + V) - S = \Delta_{V|S}$ is the residual steering gain after matched decoding; and $I_S = [(S + V) - S] - [V - G]$ is the paired difference-in-differences interaction. All effects are percentage points; confidence intervals are 95% paired-bootstrap intervals.

Model	Dataset	Method	$V - G$	$S - V$ (95% CI)	$(S + V) - S$ (95% CI)	I_S (95% CI)
1.5B	MATH500	SEAL	$+6.75$	$+2.44$ $[-1.13, 6.03]$	$+1.03$ $[-1.06, 3.22]$	-5.72 $[-10.03, -1.44]$
1.5B	MATH500	ReBalance	$+4.75$	$+4.44$ $[0.53, 8.38]$	-3.00 $[-5.50, -0.53]$	-7.75 $[-12.19, -3.47]$
1.5B	MATH500	CPS	$+11.00$	-1.81 $[-5.50, 1.84]$	$+0.31$ $[-2.00, 2.63]$	-10.69 $[-15.25, -6.34]$
1.5B	GSM8K	SEAL	$+6.75$	-1.63 $[-4.69, 1.41]$	$+0.47$ $[-2.22, 3.16]$	-6.28 $[-10.59, -2.03]$
1.5B	GSM8K	ReBalance	$+3.25$	$+1.88$ $[-1.06, 4.78]$	-0.50 $[-3.31, 2.34]$	-3.75 $[-7.41, -0.13]$
1.5B	GSM8K	CPS	$+7.75$	-2.62 $[-5.84, 0.56]$	$+2.56$ $[-0.59, 5.69]$	-5.19 $[-9.47, -0.91]$
7B	MATH500	SEAL	$+2.25$	$+0.94$ $[-0.97, 2.84]$	$+0.53$ $[-1.34, 2.41]$	-1.72 $[-4.69, 1.28]$
7B	MATH500	ReBalance	$+1.75$	$+1.44$ $[-0.47, 3.34]$	-0.56 $[-2.47, 1.31]$	-2.31 $[-5.09, 0.44]$
7B	MATH500	CPS	$+3.00$	$+0.19$ $[-1.47, 1.84]$	-0.60 $[-2.22, 1.03]$	-3.60 $[-6.22, -1.00]$
7B	GSM8K	SEAL	$+1.25$	-0.47 $[-2.44, 1.50]$	-1.09 $[-2.97, 0.78]$	-2.34 $[-5.09, 0.41]$
7B	GSM8K	ReBalance	$+1.00$	-0.22 $[-2.16, 1.72]$	$+0.19$ $[-1.66, 2.03]$	-0.81 $[-3.38, 1.75]$
7B	GSM8K	CPS	$+2.00$	-1.22 $[-2.84, 0.41]$	$+0.81$ $[-0.78, 2.41]$	-1.19 $[-3.75, 1.38]$

conditions under greedy decoding. Reported stochastic accuracies are averaged over the eight seeds. For paired accuracy contrasts, per-problem correctness is first averaged across the eight seeds and then compared with the corresponding deterministic reference on the same held-out problems.

At locations where the corresponding steering method intervenes, S replaces the activation intervention with the matched sampling policy; all other positions remain greedy. The same intervention schedule is used for $S + V$, except that the native activation intervention and its cache propagation are retained.

C.1.2 Configuration-Level Accuracy and Effect Decomposition

We separate policy-level accuracy from the effect contrasts used to test decoder-level sufficiency and competition. Table 6 reports accuracy under the four decoding conditions. Because G and V are deterministic greedy-decoding references, they are reported as single accuracies; S and $S + V$ are stochastic and are reported as mean $\pm$ standard deviation across eight independent decoding seeds.

Table 7 reports the corresponding configuration-level contrasts. $V - G$ gives the native steering gain for each configuration. $S - V$ directly compares matched stochastic decoding with native steering, while $(S + V) - S$ is the residual steering gain once matched stochastic decoding is already present. The final interaction,

$$I_S = [(S + V) - S] - [V - G], \tag{11}$$

measures how much the marginal steering gain changes after the matched sampling policy has already relaxed greedy commitment.

Table 8: **Selected temperature and top-p values and greedy-departure diagnostics.** W_1 is the Wasserstein distance between the steering and stochastic greedy-departure profiles. "dep." denotes departure from the unsteered top-1 action, and "Mean depth (V/S)" gives the mean positive departure depth for native steering and sampling, respectively.

Model	Dataset	Method	T	top-p	W_1	Steering dep.	Sampling dep.	Mean depth (V/S)
1.5B	MATH500	SEAL	0.8	0.98	0.0373	24.04%	20.83%	1.202 / 1.350
1.5B	MATH500	ReBalance	1.8	0.20	0.0201	9.47%	8.26%	1.036 / 0.985
1.5B	MATH500	CPS	1.2	0.98	0.0656	32.79%	30.10%	2.080 / 2.357
1.5B	GSM8K	SEAL	1.0	0.90	0.0229	26.85%	26.94%	1.105 / 1.155
1.5B	GSM8K	ReBalance	1.6	0.20	0.0395	9.43%	8.95%	0.964 / 0.575
1.5B	GSM8K	CPS	1.2	0.98	0.0581	36.71%	36.22%	2.071 / 2.180
7B	MATH500	SEAL	0.8	0.95	0.0182	23.74%	23.84%	1.100 / 1.023
7B	MATH500	ReBalance	0.4	1.00	0.1189	10.31%	15.08%	1.663 / 0.546
7B	MATH500	CPS	1.0	0.98	0.0535	27.40%	28.60%	1.680 / 1.620
7B	GSM8K	SEAL	0.8	0.90	0.0143	21.15%	20.22%	0.876 / 0.845
7B	GSM8K	ReBalance	2.8	0.01	0.0251	5.52%	5.68%	1.192 / 0.794
7B	GSM8K	CPS	1.0	0.98	0.0487	24.60%	25.20%	1.420 / 1.460

The two tables correspond to the two mechanistic tests in the main text. Table 6 shows that matched stochastic decoding usually achieves accuracy comparable to native steering, even though it samples variable non-greedy continuations rather than reproducing steering's fixed local choice. Table 7 shows that the residual steering gain $(S + V) - S$ is smaller than the corresponding native gain $V - G$ in every configuration, and that the paired interaction I_S is negative in all 12 rows. The strongest attenuation appears in the 1.5B MATH500 settings, while the 7B settings exhibit smaller native gains and correspondingly smaller interactions. Macro-average effects reported in the main text are computed from these within-configuration contrasts rather than from pooled raw accuracies.

C.2 Frozen-Prefix Greedy-Departure Profile Matching

The matching objective is fit only on fixed unsteered prefixes. This section reports the selected sampling parameters and frozen-prefix matching diagnostics; on-policy stability after earlier sampled departures is examined separately in Appendix C.3.

C.2.1 Selection of the Matching Parameters

Parameter selection uses only the 100-example split. For each model, dataset, and steering method, we replay the same intervention locations used by the corresponding steering method along fixed unsteered trajectories. At each site, unsteered and steered forward passes share the same textual prefix and KV cache. The steering-selected action is scored under the unsteered logits to obtain $\delta_V(s)$. For every candidate pair (T, p), the ordinary temperature-scaled top-p distribution $\pi_{T,p}(\cdot \mid s)$ induces a distribution of $\delta_S(s, a) = z_0(a_G \mid s) - z_0(a \mid s)$ at the same state. Aggregating these induced distributions over intervention sites gives $P_{T,p}(\delta)$. We search a predefined grid over temperature and top-p and select the pair that minimizes the one-dimensional Wasserstein distance in Eq. 4. Site-level contributions are averaged within problem before the 100 problems are weighted equally.

Table 8 reports the selected decoding parameters and the greedy-departure diagnostics. W_1 measures the discrepancy between the steering and sampling departure distributions. *Steering dep.* is the fraction of steering intervention sites at which the steered action differs from the unsteered top-1 action. *Sampling dep.* is the corresponding non-top-1 rate for the sampling policy, and *Mean depth (V/S)* reports the mean positive greedy-departure depth for native steering and sampling, respectively.

C.2.2 Greedy-Departure Profile Matching

The Wasserstein objective compares the complete departure-depth distributions rather than matching a single moment. For the selected (T, p), we compare $P_{T,p}(\delta)$ with $P_V(\delta)$ using their empirical CDFs. This diagnostic jointly reflects how often ordinary sampling retains the top-1 action and how deeply it moves into lower-ranked alternatives when it departs. Because top-1 retention is part of the ordinary sampling distribution, no separate departure-probability parameter is introduced. Figure 4 shows the resulting frozen-prefix CDF match for the 1.5B model on MATH500.

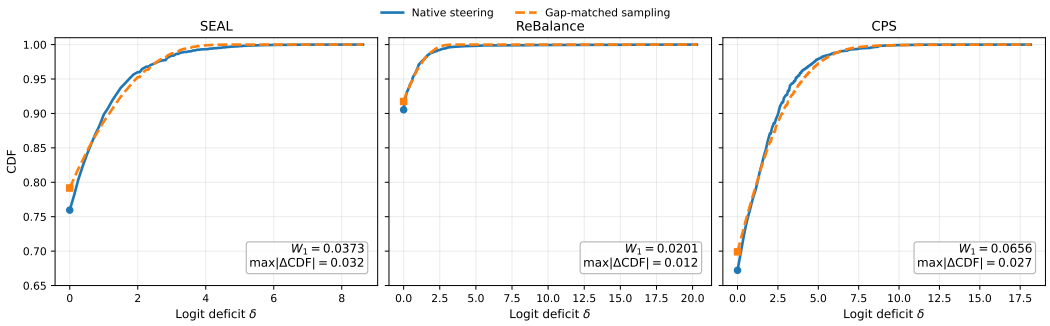

Figure 4: **Frozen-prefix greedy-departure profile matching for 1.5B on MATH500.** Empirical CDFs compare the departure depth induced by native steering with that induced by the selected standard sampling distribution at the same frozen intervention prefixes for SEAL, ReBalance, and CPS.

Table 9: **On-policy stability of matched stochastic departures for 1.5B on MATH500.** "Frozen-prefix dep." is the non-top-1 departure rate predicted on the fixed prefixes used for matching; "On-policy dep." is the realized departure rate after earlier stochastic choices have changed the trajectory. The final column reports the mean positive realized greedy-departure depth.

Method	Frozen-prefix dep.	On-policy dep.	On-policy mean depth
SEAL	20.83%	24.73%	1.351
ReBalance	8.26%	8.86%	0.985
CPS	30.10%	35.57%	2.259

C.3 On-Policy Profile Stability

The matching parameters in Appendix C.2 are selected on fixed unsteered prefixes, whereas stochastic departures change the subsequent autoregressive context distribution. We therefore re-measure departure statistics on trajectories generated under S. This on-policy analysis is reported for DeepSeek-R1-Distill-Qwen-1.5B on MATH500.

The *on-policy departure rate* is the fraction of visited intervention sites at which S selects a token different from the unsteered top-1 action. The *on-policy mean depth* is the average $z_0(a_G \mid s) - z_0(a_S \mid s)$ over those realized non-greedy selections. Table 9 compares these realized quantities with the frozen-prefix departure rate predicted during matching.

The on-policy rates remain close to the frozen-prefix predictions for ReBalance and rise moderately for SEAL and CPS. The realized mean depths remain on the same scale as the corresponding frozen-prefix sampling depths, indicating that the selected sampling policies continue to induce nontrivial but controlled departures after the state distribution shifts on-policy.

C.4 Whole-Thought Exploration Details

C.4.1 Exploration Policy and Threshold Selection

Whole-thought exploration uses the unsteered model's local uncertainty to determine where a reasoning segment is sampled. At each thought boundary, we compute

$$m(s_t) = z_{(1)}(s_t) - z_{(2)}(s_t) = \log \frac{p_{(1)}(s_t)}{p_{(2)}(s_t)}. \tag{12}$$

Smaller values indicate a less decisive top-1 choice. On an independent parameter-selection split, we set

$$\theta_{\mathrm{q10}} = Q_{0.10}(m), \qquad \theta_{\mathrm{q25}} = Q_{0.25}(m), \qquad \theta_{\mathrm{q50}} = Q_{0.50}(m). \tag{13}$$

Table 10: **Margin thresholds for whole-thought exploration.** Values are top-1/top-2 logit margins of the unsteered model. The thresholds are empirical quantiles of the corresponding parameter-selection thought-boundary margin distribution.

Dataset	Model	θ_{q10}	θ_{q25}	θ_{q50}
MATH500	1.5B	0.250	0.875	2.375
MATH500	7B	0.250	0.750	1.750
GSM8K	1.5B	0.375	1.125	3.000
GSM8K	7B	0.250	0.750	2.000

For dose $k \in \{\mathrm{q10}, \mathrm{q25}, \mathrm{q50}\}$, whenever $m(s_t) \leq \theta_k$, the next complete reasoning segment is generated by sampling until the next thought boundary; otherwise the segment is generated greedily. The margin is then recomputed at the new boundary. Under ALL, sampled-segment generation is applied at every thought boundary. Accordingly, q10, q25, and q50 target approximately the lowest 10%, 25%, and 50% of the parameter-selection margin distribution, respectively. Table 10 reports the resulting model–dataset-specific thresholds.

Because sampled segments alter subsequent trajectories, the fraction of thought boundaries that actually trigger exploration need not equal the nominal quantile exactly; the quantile specifies the decision rule rather than a fixed realized sampling budget. Ties at a quantile threshold can likewise make the realized trigger fraction differ slightly from the nominal percentage.

C.4.2 Evaluation and Aggregation Protocol

We evaluate the same model, dataset, and steering-method settings as in the main greedy-departure experiment. For each dose, E_k uses the unsteered margin to determine whether the next thought is sampled, while $E_k + V$ applies the same trigger rule together with native steering. All effects are first computed within each configuration and then macro-averaged with equal configuration weight. Confidence intervals use 20,000 paired problem-level bootstrap replicates; within each dataset, the same resampled problem indices are shared across methods to preserve correlations induced by common greedy and exploration outputs. The aggregate dose-response and competition results are reported in Table 3 in the main text.

C.4.3 Configuration-Level Residual Gain and Interaction

The aggregate results summarize the overall competition pattern, but the magnitude of the native and residual steering effects differs across model–dataset–method configurations. Figure 5 gives a compact view of the configuration-level residual gains together with the equal-weight macro-average.

Figure 6 expands the same residual effect into one panel per configuration. Error bars are configuration-level 95% paired-bootstrap confidence intervals, and the dashed horizontal line in each panel marks the corresponding native steering gain $\Delta_V = A(V) - A(G)$. At q10, several 1.5B configurations retain substantial positive residual steering gains. From q25 onward, the residual effect generally contracts toward zero and is much smaller than the native steering reference, although the exact dose response varies across configurations. The 7B settings begin with smaller native steering gains overall and correspondingly show smaller residual effects throughout.

A complementary view normalizes each configuration by its own native steering unsteered greedy condition through the interaction

$$I_{E_k} = [A(E_k + V) - A(E_k)] - [A(V) - A(G)] = \Delta_{V|E_k} - \Delta_V. \tag{14}$$

Negative values indicate that whole-thought exploration reduces the marginal value of steering relative to the unsteered greedy condition. Figure 7 reports this quantity for the same 12 configurations. The strongest negative interactions appear mainly in the 1.5B settings from q25 onward, while the 7B settings are typically closer to zero because their native steering gains are smaller to begin with. Together, Figures 6 and 7 show the same pattern from two angles: moderate-to-strong exploration both shrinks the residual gain left for steering and reduces that gain relative to steering's native effect under greedy decoding.

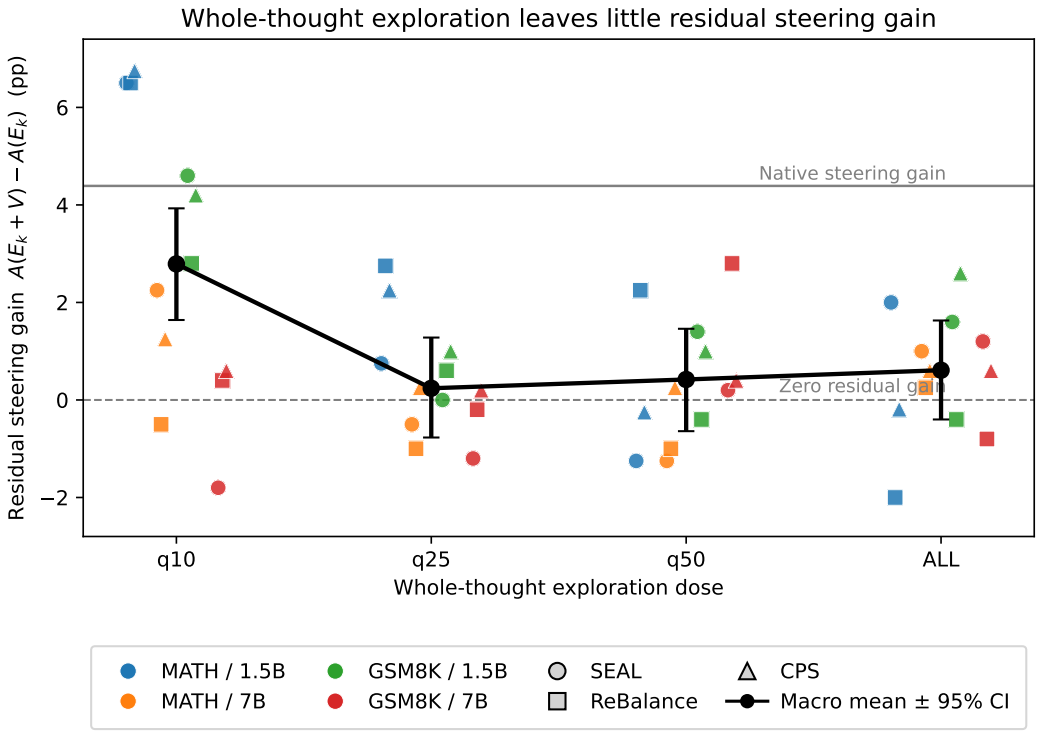

Figure 5: **Whole-thought exploration reduces steering's marginal gain.** Small points show configuration-level residual steering gains $\Delta_{V|E_k} = A(E_k + V) - A(E_k)$. The larger points and error bars show the equal-weight macro-average and its 95% paired-bootstrap confidence interval at each exploration dose. The solid horizontal reference marks the native steering gain under greedy decoding, and the dashed line marks zero residual steering gain.

D Additional Methodological Details

D.1 Steering Construction and Data Separation

For each base model, steering construction uses a model-specific MATH response pool containing 500 correctly answered and 500 incorrectly answered trajectories. These construction problems are disjoint from the MATH500 and GSM8K evaluation examples. The same 1,000-response construction pool is supplied to SEAL, ReBalance, and CPS for a given base model. SEAL and ReBalance use their official construction and inference implementations. CPS uses the cluster-prototype construction described below. No steering direction is reselected using the outcomes of the mechanism experiments reported in the paper.

D.2 CPS Construction and Inference

Trajectory and representation collection. For the 1.5B CPS direction used in the detailed mechanism analyses, we first generate complete mathematical reasoning trajectories with DeepSeek-R1-Distill-Qwen-1.5B on MATH training problems, with a maximum of 10,000 new tokens per generation. The representation corpus contains 500 trajectories with correct final answers and 500 with incorrect final answers. Correctness is used only to balance the source trajectories; it does not define positive and negative endpoints of the steering direction.

Within each trajectory, we retain the `<think>...</think>` reasoning region and use the consecutive-newline thought-split token as the reasoning boundary. At the t-th boundary of tra-

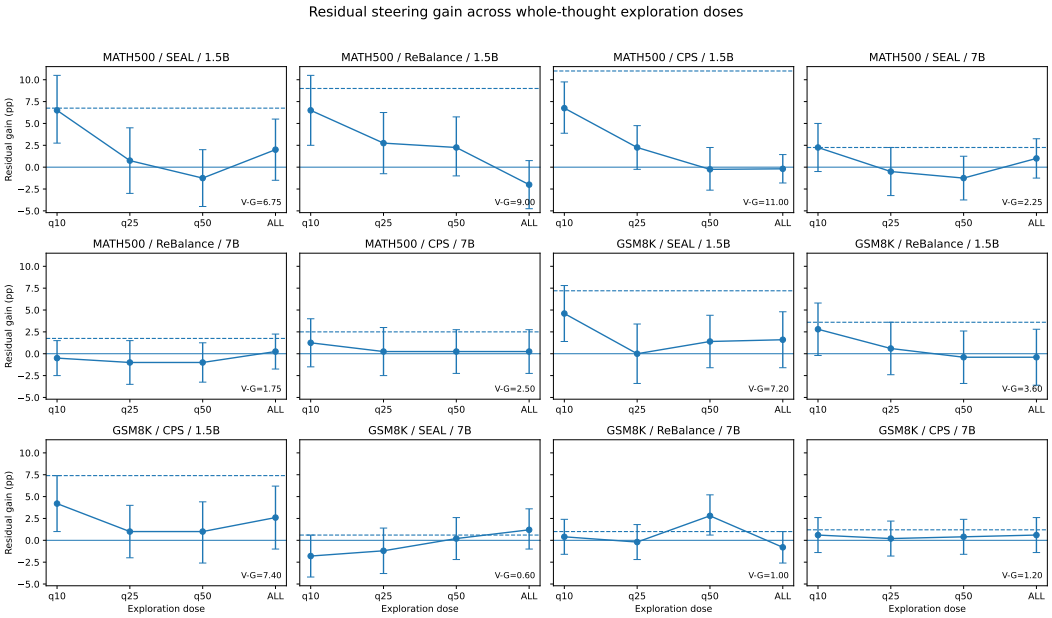

Figure 6: **Configuration-level residual steering gain across whole-thought exploration doses.** Each panel corresponds to one dataset–method–model configuration. Points and error bars show $\Delta_{V|E_k} = A(E_k + V) - A(E_k)$ at q10, q25, q50, and ALL, with 95% paired-bootstrap confidence intervals. The dashed horizontal line marks the native steering gain $\Delta_V = A(V) - A(G)$ for the same configuration; it is independent of the exploration dose.

jectory i, let

$$h_{i,t}^{(20)} \in \mathbb{R}^{1536} \tag{15}$$

denote the residual-stream representation immediately before Transformer block 20. Layer numbering is zero-based, so this representation follows the first 20 blocks.

Unsupervised prototype construction. We project the collected boundary representations to 20 principal components,

$$x_{i,t} = \mathrm{PCA}_{20}\left(h_{i,t}^{(20)}\right), \tag{16}$$

and apply K-means with six clusters. The PCA space is used only for clustering. For each cluster $\mathcal{C}_k$, we return to the original 1536-dimensional residual space and compute

$$\mu_k = \frac{1}{|\mathcal{C}_k|} \sum_{(i,t)\in\mathcal{C}_k} h_{i,t}^{(20)}, \qquad k = 1, \ldots, 6. \tag{17}$$

The resulting prototypes are data-driven boundary representations; we do not assign them prespecified labels such as reflection, execution, or correctness.

Candidate direction and selection. Candidate CPS directions are linear combinations of the six original-space cluster means,

$$v(\boldsymbol{\beta}) = \sum_{k=1}^{6} \beta_k \mu_k. \tag{18}$$

The original CPS search protocol compares candidate combinations and local coefficient refinements and then fixes one direction $v_{\mathrm{CPS}} = v(\boldsymbol{\beta}^\star)$ for subsequent analysis. The selected 1.5B direction has dimension 1536 and L_2 norm 115.12 and is used without unit-norm normalization. The direction and its coefficients remain fixed throughout the mechanism experiments.

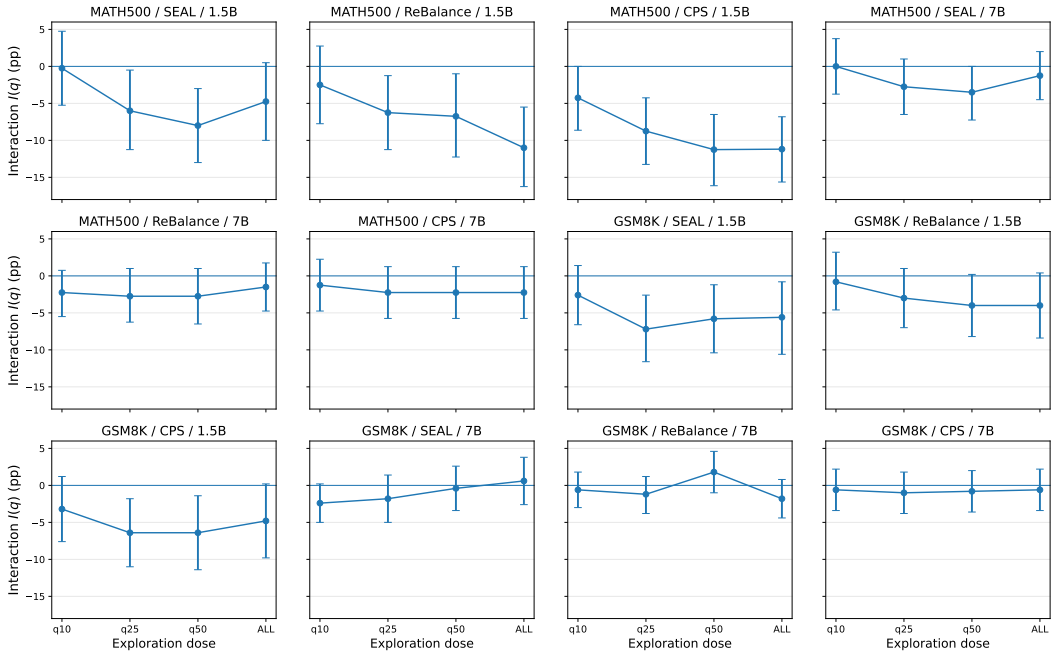

Figure 7: **Configuration-level interaction across whole-thought exploration doses.** Each panel corresponds to one dataset–method–model configuration. Points and error bars show $I_{E_k} = \Delta_{V|E_k} - \Delta_V$ at q10, q25, q50, and ALL, with 95% paired-bootstrap confidence intervals. Negative values indicate that whole-thought exploration weakens the marginal accuracy gain of steering relative to its native effect under greedy decoding.

CPS should therefore be interpreted as a preselected analysis intervention rather than as a newly tuned estimator on each reported experiment. Its original MATH500 accuracy is used to characterize the selected intervention, while subsequent mechanism analyses keep the direction fixed and do not reselect it from their test outcomes.

Online intervention. CPS is applied only at reasoning boundaries inside the `<think>...</think>` region. If $r_{\ell,t}$ denotes the residual stream entering block ℓ at position t, then at an eligible boundary we apply

$$\widetilde{r}_{20,t} = r_{20,t} + \alpha v_{\mathrm{CPS}}, \tag{19}$$

with $\alpha = 1.0$ in the main experiments. The modified representation then proceeds through block 20 and all subsequent blocks. No CPS intervention is applied to the problem prompt, ordinary reasoning tokens, or the answer text after `</think>`. The same fixed vector and scale are applied at every eligible boundary, while the resulting logit change remains context dependent because the surrounding generation context changes over time.

Cache and generation behavior. Generation uses the standard autoregressive KV cache. Each intervention directly modifies only the current boundary representation; it does not retroactively alter cached representations for earlier tokens. Its downstream keys and values can nevertheless enter the cache and affect later generation. The default CPS condition uses greedy decoding (`do_sample=False`) with a 10,000-token generation limit. Generation does not stop at the first boxed answer or at `</think>`; it ends when the model emits EOS or reaches the length limit.