CONFLICT STEERING: RESOLVING LATE-LAYER DISAGREEMENT REPAIRS NON-TERMINATION IN GREEDY REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning models are shipped with a warning not to decode them greedily: on a long trace a greedy decoder resolves every recurring near-tie in the same way, so it can loop or wander to the token limit. Sampling breaks such ties at random, but the output then depends on the draw. We show that the model's own late layers supply a deterministic tie-breaker: within a decodable band of the last M layers, an intermediate layer sometimes commits to a token with low entropy and the final layer overrides it. We call these steps *fights* and resolve them by *conflict steering*, a spherical interpolation of the final hidden state toward the intermediate one, followed by ordinary hard-token decoding. Conflict steering repairs a failure mode, not a capability: the non-termination of a derivation the model can complete. Over four Qwen3 models spanning 4B to 32B one steered greedy decode wins 189 paired problems and loses 141 against greedy on AIME 2024 to 2026 and three code benchmarks. The gain sits where greedy leaves traces open: at 4B and 8B, where a fifth of AIME traces reach the token limit, it wins 112 and loses 70 and surpasses the accuracy of temperature-0.6 sampling on AIME in three of four serving configurations. On multiple-choice, near-ceiling and beyond-competence controls it predicts no gain and finds none. At 14B and 32B, where under a tenth of the traces are open, one decode gains little on code and, at 14B, stays below the sampled mean on AIME. The steering fallback, conflict steering only on non-terminating greedy traces, never loses in scope (92 wins, no losses over the four models). Under sampling, whose randomness already breaks ties, no training-free latent-reasoning method we evaluate, ours included, improves on plain sampling.

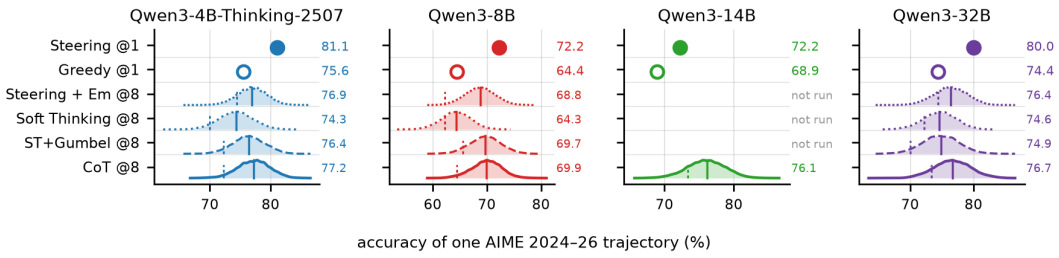

Figure 1: One deterministic decode against one temperature-0.6 sample on AIME 2024–26, four Qwen3 models. Ridges: accuracy of one sample per problem, bootstrapped from eight runs (solid tick and label: mean, dotted tick: minimum run), markers: one greedy (open) and one conflict-steered (filled) decode. Steering beats greedy for every model and the sampled mean at 4B, 8B and 32B, at 14B it is 3.9 below.

1 INTRODUCTION

Reasoning models are shipped with recommended sampling temperatures rather than greedy decoding: Qwen3 warns that greedy reasoning loses accuracy and repeats indefinitely (Yang et al., 2025), and open reasoning models across families loop at low temperature (Pipis et al., 2026). On a long

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.

reasoning trace, a greedy decoder resolves every recurring near-tie the same way, so once in a cycle it stays there until the token limit. Sampling breaks such ties at random, but the accuracy of a single draw varies (Figure 1), and a more reliable answer needs several draws and a vote or verifier. Reproducible evaluation, audited pipelines, and cost-constrained serving instead need one deterministic trajectory (Gond et al., 2026; Yuan et al., 2026; Aggarwal et al., 2023; Snell et al., 2025). We ask whether information already present inside the model can give that trajectory a better tie-breaker than the argmax.

Late-layer disagreement. Confident Layer Decoding (CLD) (Zhang et al., 2026b) decodes the last M layers with the model's own final norm and unembedding and, when the layer entropies form a trough, emits the top token of the trough layer. We instead use the trough as a conflict signal: if its top token differs from the final layer's, an earlier low-entropy choice has been overridden late in the network, an event we call a *fight*.

Resolving the disagreement. Committing the intermediate token at every trough, as CLD does, does not help consistently. Conflict steering instead moves the final hidden state toward the intermediate state by spherical interpolation (You et al., 2026) only at fights. The final norm and unembedding then produce an ordinary hard token, so the token stream and key-value cache are unchanged.

Repairing a failure mode, not adding capability. Conflict steering repairs one failure of greedy decoding, the non-termination of a derivation the model can complete, and the mechanism says where it should help: where greedy leaves many traces at the token limit on tasks with a generated, verifiable answer. It should also do nothing on near-ceiling benchmarks or on problems beyond the model's competence. We treat these as controls stated in advance. We test four Qwen3 models spanning 4B to 32B, including an 8B-to-32B scale sweep, on AIME and code. Our experiments support the prediction: at 4B and 8B, the gain comes mainly from traces that greedy decoding leaves open, and a single steered decode reaches or exceeds the expected accuracy of temperature-0.6 sampling on AIME. Along the scale sweep, the termination gain shrinks together with the open-trace rate, while the controls show no such gain (see Section 5.1 for details)

Contributions.

- **A deterministic tie-breaker from late-layer conflict.** We identify low-entropy intermediate decisions that the final layer overrides and introduce conflict steering, a training-free, JSD-controlled spherical interpolation that turns this disagreement into an ordinary hard token. It requires no verifier or labeled deployment calibration and leaves the token stream and key-value cache unchanged.
- **One deterministic decode can match sampling where greedy fails to terminate.** On AIME and code at Qwen3 4B and 8B, conflict steering wins 112 paired problems and loses 70, reaches or exceeds the temperature-0.6 sampled mean on AIME, and improves every code benchmark at both scales.
- **A predictive boundary and a conservative deployment rule.** Trace-state accounting shows that most of the 4B and 8B in-scope gain comes from traces greedy leaves open. Three mechanism controls show no drop-in gain, and along the 8B to 32B sweep the termination gain shrinks with the open-trace rate. Rerunning only open non-terminating greedy traces with steering yields 92 wins and no losses in scope over all four models, requiring a second decode for only 3–20% of problems, corresponding roughly to 1.06–1.40$\times$ greedy compute, rather than 2$\times$ for always-on steering or roughly 8$\times$ for eight sampled trajectories.

2 Background and related work

Notation. A decoder-only transformer with L blocks produces a hidden state $\boldsymbol{h}^{(\ell)} \in \mathbb{R}^D$ after block ℓ at each decoding step. We write $F = L$ for the final layer. Applying the final normalization and the unembedding $\boldsymbol{W} \in \mathbb{R}^{V \times D}$ (the output projection, or language-model head) gives logits $\boldsymbol{z}^{(\ell)} = \boldsymbol{W}\,\mathrm{Norm}(\boldsymbol{h}^{(\ell)})$ and distribution $p^{(\ell)} = \mathrm{softmax}(\boldsymbol{z}^{(\ell)})$. Reading $p^{(\ell)}$ before the final layer is the logit lens (nostalgebraist, 2020; Belrose et al., 2023; Ghandeharioun et al., 2024). We use $H(p)$

for Shannon entropy in nats and $\mathrm{JSDn}(p^{(I)}, p^{(F)}) \in [0, 1]$ for the Jensen–Shannon divergence in base 2 between an I-th intermediate layer's distribution and the final layer's.

Greedy decoding and sampling. Greedy decoding can enter self-reinforcing repetition loops (Holtzman et al., 2020; Fu et al., 2021; Xu et al., 2022) while sampling and presence penalties are prescribed to break them (Yang et al., 2025). Higher-temperature sampling can break low-temperature loops caused by self-reinforcing choices or errors (Pipis et al., 2026), but it does so by chance rather than correcting the underlying failure. AIME traces are shown to be bimodal, converging early or never halting (Oladri et al., 2026). Spending more test-time compute can improve on plain sampling by self-consistency (Wang et al., 2022), adaptive stopping (Aggarwal et al., 2023), confidence-filtered parallel thinking (Fu et al., 2026), or power sampling from the base likelihood (Karan & Du, 2026). Reshaping the sampling distribution by min-p (Nguyen et al., 2025), entropy-adaptive temperature (Zhang et al., 2024c), adaptive foresight sampling (Xu et al., 2025), or adaptive top-k decoding (Noarov et al., 2026) can also improve sampling. All of these presuppose sampled traces, whereas ours allows one deterministic decode.

Training-free latent reasoning. Soft Thinking (Zhang et al., 2025b) feeds back a probability-weighted mixture of token embeddings, then returns to hard tokens after a low-entropy *cold stop*. Wu et al. (2026) add Gumbel noise to the mixture. Multiplex Thinking (Tang et al., 2026) trains on token-wise branch-and-merge inputs. SwiReasoning (Shi et al., 2026) uses entropy-gated alternates soft and hard reasoning blocks. SeLaR (Fu & Luo, 2026) inserts a soft embedding only at high-entropy steps. Trained latent-reasoning models feed hidden states back as continuous thought input (Hao et al., 2025) or recurs in depth by repeatedly applying a middle block (Park et al., 2026). Instead, we read conflict directly from the model's interpretable layers while keeping the generated sequence in the ordinary token space.

Layer-wise decoding and steering. Early-exit and layer-skipping methods decode from a shallower intermediate layer to save compute or to draft tokens that the full model then verifies (Schuster et al., 2022; Elhoushi et al., 2024; Zhang et al., 2024b; Entezari Zarch et al., 2025; Wei et al., 2026; Ma et al., 2026). DoLa (Chuang et al., 2024), its conditional variant (Zhang et al., 2025a), SLED (Zhang et al., 2024a), and DeLTa (He et al., 2025) contrast intermediate and final logits to improve or amplify factual predictions, and inter-layer disagreement can detect hallucinations (Subbalakshmi et al., 2026), attribution graphs verify and repair reasoning steps (Zhao et al., 2026), and CLD (Zhang et al., 2026b) commits the top token of an entropy-trough layer. Spherical steering (You et al., 2026) rotates activations toward a direction estimated from example pairs, motivated by evidence that angle carries meaning while norm governs stability or intensity (Aparin & Gaintseva, 2026; Loshchilov et al., 2025; Park et al., 2024; Zhang et al., 2026a). We use the intermediate state itself as the target, without example pairs, only when its top token disagrees with the final layer's.

3 METHOD: DETECTING AND STEERING CONFLICT

Conflict steering has three parts: a band of decodable late layers, a detector for disagreement within that band, and a spherical interpolation that resolves the disagreement. All three are training-free and applied at each decoding step (Figure 2).

3.1 THE DECODABLE BAND

We decode the last M layers, $\ell \in \{F - M + 1, \ldots, F\}$, with the pre-trained final norm and unembedding. This is useful only where the intermediate distributions are already close to the final one: their mean normalized JSD falls below 0.1 in the last few layers, and layer $F-1$ is a near-copy of F at every scale ($\mathrm{JSDn} < 0.005$), so the band needs at least three layers for disagreement to occur.

3.2 FIGHTS: THE CONFLICT DETECTOR

At each step, we compute $p^{(\ell)}$ and entropy H_ℓ for every layer in the band. Following CLD, we find an intermediate layer $I < F$ whose entropy is lower than its neighbors', a local trough. Let

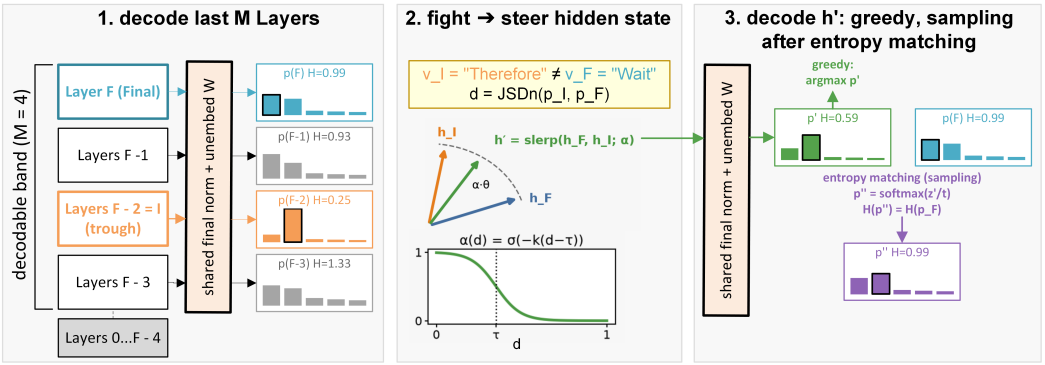

Figure 2: One conflict-steering step. Left: the last M layers decoded through the shared final norm and unembedding. The trough layer I (orange) has the band's lowest entropy and its top token differs from the final layer's (blue): a fight. Middle: the final hidden state moves along the great circle toward the trough state, norm kept, by $\alpha(d)$ (Equation 2). Right: $\boldsymbol{h}'$ decoded through the same head, greedy takes the argmax of p', sampling first rescales the logits until $H(p'') = H(p_F)$, keeping the ranking. Bars: illustrative distributions with 8B mean entropies.

$v_I = \arg\max_v p^{(I)}(v)$ and let v_F be the final layer's top token. A *fight* occurs when a trough exists and $v_I \neq v_F$. All other steps are decoded normally. The median fight rate is 3.1% to 8.5% per generated token across the 4B and 8B AIME and GPQA traces.

We measure the disagreement as $d = \mathrm{JSDn}(p^{(I)}, p^{(F)}) \in [0, 1]$: small d means the layers disagree on the top token but agree on the rest of the vocabulary, and large d a broader disagreement.

3.3 Conflict steering

CLD resolves a fight by emitting v_I. We instead move the final hidden state toward the intermediate state, then apply the final norm and unembedding. Let θ be the angle between $\boldsymbol{h}_F$ and $\boldsymbol{h}_I$. Spherical linear interpolation by a fraction $\alpha \in [0, 1]$,

$$\mathrm{slerp}(\boldsymbol{h}_F, \boldsymbol{h}_I; \alpha) = \frac{\sin((1-\alpha)\theta)}{\sin\theta}\,\boldsymbol{h}_F + \frac{\sin(\alpha\theta)}{\sin\theta}\,\boldsymbol{h}_I, \tag{1}$$

moves along the geodesic path from $\boldsymbol{h}_F$ to $\boldsymbol{h}_I$ while keeping the norm on the scale of the endpoints. The angle is small in practice (median 24° at 8B). We set the interpolation fraction with a decreasing function of the disagreement:

$$\alpha(d) = \lambda_{\mathrm{far}} + (\lambda_{\mathrm{near}} - \lambda_{\mathrm{far}})\,\sigma\big(-k\,(d-\tau)\big), \qquad \sigma(x) = \frac{1}{1+e^{-x}}, \tag{2}$$

Here τ is the threshold, k the sharpness of the transition, and λ_{near}, λ_{far} the interpolation at small and large disagreement. We use $\lambda_{\mathrm{near}} = 1$, $\lambda_{\mathrm{far}} = 0$, $k = 14$ and $\tau = 0.35$ throughout, so the fraction falls from 0.9 to 0.1 over about 0.3 of JSDn around τ. This means a fight confined to the top token receives a strong interpolation, while broad disagreement receives little or none. We decode $\boldsymbol{h}' = \mathrm{slerp}(\boldsymbol{h}_F, \boldsymbol{h}_I; \alpha(d))$ as $p' = \mathrm{softmax}(\boldsymbol{W}\,\mathrm{Norm}(\boldsymbol{h}'))$. The next step receives the selected hard token and the key-value cache is unchanged. At 8B on AIME a strong interpolation selects v_I in 73% of fights and keeps v_F in 25%, so steering does not simply commit the intermediate token.

3.4 Steering under sampling: entropy matching

Greedy decoding depends only on the argmax of p' while sampling also depends on the mass left for alternatives. Because the trough is the entropy minimum of the band (0.11 versus 0.46 nats at 8B), interpolation usually makes p' sharper than $p^{(F)}$, which under sampling produces more repetition loops (Figure 3. For sampling, we therefore choose $t > 0$ using bisection with $H(\mathrm{softmax}(\boldsymbol{z}'/t)) = H(p^{(F)})$ on the sampler's re-normalized top-k support (variant Em) and decode $\boldsymbol{z}'/t$. A positive temperature preserves the ranking, so entropy-matched steering gives exactly the greedy trace of plain steering.

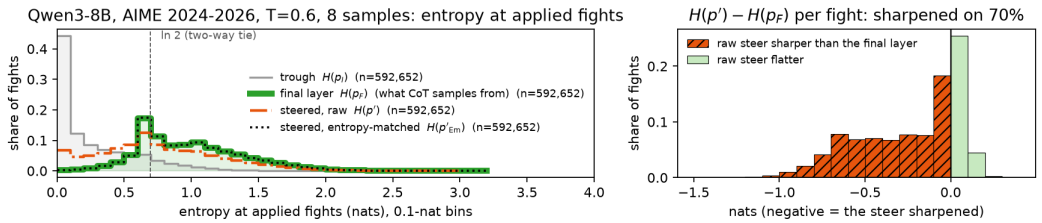

Figure 3: Entropy at applied fights (Qwen3-8B, AIME 2024–26, temperature 0.6, eight samples). Left: histograms of the trough (thin gray), final-layer (thick green), raw steered (orange dash-dot) and entropy-matched (black dotted) distributions, line at $\ln 2$. Right: raw-steered minus final-layer entropy per fight. Hatched marks negative values, where steering sharpens the distribution (70%).

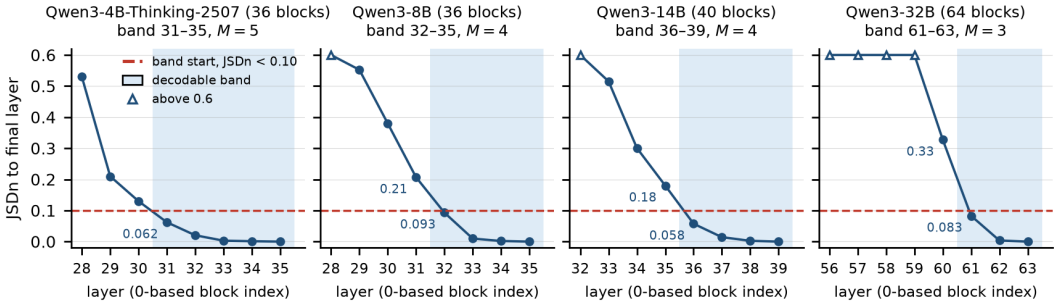

Figure 4: Layer probe on AIME 2024 (30 problems at 8B and 14B, 10 at 4B and 32B): $\overline{\text{JSDn}}$ of each late layer to the final layer. Dashed: band start at 0.10, shaded: selected band, triangles: layers above 0.6.

4 Calibration from an unlabeled probe

The band width M and threshold τ are read from a short unlabeled probe of a few hundred decoding steps, and the remaining parameters are fixed at $\lambda_{\text{near}} = 1$, $\lambda_{\text{far}} = 0$ and $k = 14$. The probe does not use answer accuracy. The rule itself was developed on the first 4B and 8B draws and then frozen for 14B, 32B and the distilled model of Appendix B.2, so the deployed calibration is unlabeled while its development was not blind to the 4B and 8B results.

Band width. For each layer, we record its normalized JSD to the final layer (the median over decode steps of each probe problem, averaged over problems, written $\overline{\text{JSDn}}$) and its mean top-1 agreement with the final layer. The profile has the same form for all four models. Intermediate distributions are far from the final one until a cliff spanning several layers where the distance drops to a short group of close layers that still sometimes choose a different top token.

The band starts at the first layer whose $\overline{\text{JSDn}}$ to the final layer falls below 0.10 (Figure 4). At every model we probed this leaves at least three layers (layer $F - 1$ is a near-copy of F, so $M \geq 3$). This gives $M = 5$, 4, 4 and 3 at 4B, 8B, 14B and 32B (start layers 31, 32, 36 and 61).

Disagreement threshold. Across all trough steps, d has one mode near 0.07 and another near 0.5, with a gap around 0.35, so we set $\tau = 0.35$, but among fights alone the distribution is unimodal and similar across contest math, science and code. About 30 percent of 8B fights and 41 percent of 4B fights receive the full move. Appendix A.1 reports in-depth analysis on both M and τ.

5 Experiments

Setup. All models run in bfloat16 on A100-80GB GPUs (14B and 32B with tensor parallel 2) on our SGLang 0.4.6 fork of Soft Thinking (Zhang et al., 2025b; Zheng et al., 2024). We evaluate four Qwen3 models (Yang et al., 2025): a three-scale sweep, Qwen3-8B, Qwen3-14B and Qwen3-32B,

Table 1: Accuracy in percent on AIME 2024–26 (90 problems) and the pooled code sets (700) for four Qwen3 models. Sampled: mean ± standard deviation of the per-run accuracy over eight chain-of-thought runs at temperature 0.6, and the minimum run (not run on code at 32B). Deterministic: one decode at a fixed batch size. Steering fallback: greedy, then one steered rerun of each trace that hit the token budget. Shaded: ours. Bold: best per column. Underlined: best single decode where a sampled mean is higher.

	Qwen3-4B-Thinking-2507 ($M = 5$)		Qwen3-8B ($M = 4$)		Qwen3-14B ($M = 4$)		Qwen3-32B ($M = 3$)	
Method	AIME-3yr	Code (700)	AIME-3yr	Code (700)	AIME-3yr	Code (700)	AIME-3yr	Code (700)
Sampled, temperature 0.6, eight runs								
CoT @8, mean ± sd	77.2 ± 3.2	**83.6** ± 0.6	69.9 ± 3.0	**81.1** ± 0.6	**76.1** ± 2.7	83.6 ± 0.9	76.7 ± 3.1	—
CoT @8, minimum run	72.2	82.3	64.4	80.0	73.3	82.3	73.3	—
Deterministic								
Greedy	75.6	81.0	64.4	74.4	68.9	81.3	74.4	84.9
CLD hard (commit v_I)	75.6	82.0	63.3	73.9	—	—	—	—
Deterministic conflict steering								
Conflict steering	81.1	<u>83.0</u>	**72.2**	76.6	72.2	82.1	**80.0**	83.4
Steering fallback	**83.3**	82.3	71.1	<u>79.9</u>	<u>73.3</u>	**84.1**	78.9	**85.4**

and the later thinking-only release Qwen3-4B-Thinking-2507, written 8B, 14B, 32B and 4B. The 4B model outscores the 8B model on AIME and is treated as an additional model, not the smallest point of the sweep. Batch size is fixed per model and dataset. Refer to Appendix H for detailed serving and model configuration by scale and dataset.

Generations are limited to 32,768 tokens including thinking. Greedy cells use temperature 0, and for conflict steering the calibrated band and constants in Section 3. Sampled cells are marked @8: eight samples per problem at temperature 0.6, top-p 0.95 and top-k 20, the released Soft Thinking protocol, shared by all sampled methods. We report the mean, the standard deviation and the minimum of the eight chain-of-thought and other sampling runs.

In-scope benchmarks have a generated, verifiable answer: AIME 2024, 2025 and 2026 (30 problems each, together the 90-problem AIME-3yr) and HumanEval, MBPP and LiveCodeBench v5 (Chen et al., 2021; Austin et al., 2021; Jain et al., 2025). Three groups serve as controls: GPQA-Diamond (198, four-way multiple choice) (Rein et al., 2024); AMC23, MATH-500 and GSM8K (Lightman et al., 2024; Cobbe et al., 2021), on which greedy already scores 92 to 100 percent, and five 2025 contest sets beyond the models' competence (HMMT, BRUMO, CMIMC, SMT and BeyondAIME; 253 problems, greedy 36 to 73 percent) (Balunovic et al., 2025). We score every benchmark by rule, reading the text after `</think>` or, for an open trace, the unfinished text. Results give accuracy and the percentage of open traces (no-`</think>`). Baselines are greedy at the same batch size, CLD hard (Zhang et al., 2026b), a presence penalty, and sampled chain-of-thought, Soft Thinking and ST+Gumbel. For paired comparisons, p is the two-sided exact sign-test p-value over problems solved by exactly one method (Appendix C).

5.1 Generated-answer tasks

Conflict steering gives the highest single-decode AIME-3yr score in Table 1. The ablation, committing the token at every trough step (CLD hard), does not improve AIME and is below steering on every code set on both 4B and 8B scales. Steering improves code on all three sets at both scales: 79 wins and 50 losses over 1,400 paired problems ($p = 0.01$). The largest gain is LiveCodeBench at 8B, where greedy leaves 28 percent of traces open. See Table 9 in Appendix B.3 for individually scored code benchmarks. The standard repair, a presence penalty, requires a strength search per model (1.0 helps 8B and hurts 4B, where 0.5 helps instead); see Table 13 in Appendix E.1. Chosen on the test labels it matches steering on AIME (137 versus 138), and the steering fallback is ahead of it at 8B.

Single-draw reliability. A sampled decoder returns one trajectory per query and, in most deployments, no verifier selects among draws, so the accuracy a user receives is a random variable. Average@8 estimates its expectation and hides the spread: the per-run standard deviation is 2.7 to 3.2 points on AIME and 0.6 to 0.9 on code (the ± in Table 1). Table 1 therefore also reports the minimum of the eight runs, the observed lower tail of one draw rather than an estimate of the distributional worst case. A deterministic decoder instead yields one trajectory for a fixed serving

Table 2: AIME 2024–26 under the shared sampler (temperature 0.6, top-p 0.95, top-k 20, eight samples): Average@8 $\pm$ standard deviation of the per-run accuracy, maj@8 (vote over the eight extracted answers), truncated samples of 720 (loops in parentheses) and pooled code Average@8. Shaded: ours. Bold: best per column.

	AIME-3yr, Qwen3-4B-Thinking-2507			AIME-3yr, Qwen3-8B			Code (700)	
Method	Avg@8	maj@8	truncated (loops)	Avg@8	maj@8	truncated (loops)	4B	8B
CoT @8	77.2 ± 3.2	**86.7**	108 (0)	**69.9** ± 3.0	76.7	54 (4)	**83.6** ± 0.6	**81.1** ± 0.6
Soft Thinking @8	74.3 ± 3.3	84.4	91 (0)	64.3 ± 1.9	74.4	40 (0)	—	—
ST+Gumbel @8	76.4 ± 3.2	85.6	84 (0)	69.7 ± 2.7	**78.9**	44 (0)	—	—
Conflict steering @8	**77.9** ± 2.3	**86.7**	110 (0)	68.2 ± 2.5	76.7	98 (35)	—	—
Conflict steering, Em @8	76.9 ± 1.9	85.6	111 (0)	68.8 ± 3.4	77.8	74 (21)	82.7 ± 0.6	79.4 ± 0.7

Table 3: Controls under greedy decoding: accuracy in percent, problems won↔lost by steering against greedy, and the net gain of the steering fallback over greedy. Best per scale in bold, shaded columns ours. At 14B only GPQA was run.

		4B			8B			14B		
Control	Benchmark (n)	Greedy	Steering	Fallback	Greedy	Steering	Fallback	Greedy	Steering	Fallback
Multiple choice	GPQA-Diamond (198)	**66.7**	64.6 (10↔14)	**66.7** (+0)	59.1	56.6 (20↔25)	**59.6** (+1)	60.1	56.6 (13↔20)	**60.6** (+1)
Near ceiling	AMC23, MATH-500, GSM8K (1,859)	**96.4**	96.0 (11↔20)	**96.4** (+0)	95.6	95.1 (24↔34)	**96.0** (+8)	—	—	—
Beyond competence	Five contest sets (253)	56.9	56.5 (17↔18)	**59.7** (+7)	48.6	48.2 (22↔23)	**52.2** (+9)	—	—	—

configuration. One steered decode is above the minimum run by 8.9 and 7.8 points at 4B and 8B. On code it lies inside the two-point span of the runs at 4B and below every run at 8B, where the steering fallback narrows the gap.

Determinism is conditional on the serving configuration. Greedy decoding is deterministic for a fixed execution configuration. A change of batch size alters the floating-point reduction order, can flip an early near-tie token (Yuan et al., 2026) and thereby selects a different deterministic trajectory (at 8B, batch sizes 12 and 16 share no identical AIME completion and solve 58 and 63 of 90). This sensitivity belongs to deterministic decoding in general, and we treat it as a design constraint rather than a knob. Every comparison pairs steering with greedy at the same batch size, software state and code state, batch size is fixed per model and dataset, the main tables report the first paired draw (the hardware optimized configuration) of every cell. Appendix C tests serving-stack sensitivity across repeated problems (Table 11). Across four AIME configurations steering is ahead in three and behind by two problems in one, 33 wins and 20 losses over 360 pairs ($p = 0.10$), and with the code cells 112 versus 70 over 1,760 pairs ($p = 0.002$).

5.2 Comparison with sampling

Table 1 compares one deterministic decode with plain sampled decodes. On 8B code, steering falls below to the sampled mean by 4.5 points (76.6 versus 81.1), where sampling closes most of the traces greedy leaves open. The Steering fallback narrows the gap to just 1.2 points (79.9 versus 81.1).

Under the sampler, steering is a test under sampling, not a second method. Sampling already breaks the near-ties it resolves, so no gain is expected, and Section 3.4 predicts more repetition without entropy matching. Table 2 verifies both. No training-free latent-reasoning method improves on chain-of-thought: Soft Thinking is 5.6 points below at 8B (24↔10, $p = 0.024$), ST+Gumbel and steering are within noise at 4B and 8B, and entropy-matched steering is 0.3 below at 32B (Table 12 of Appendix D.1). Plain steering sharpens the distribution and truncates 98 samples against 54 at 8B (35 loops against 4). Entropy matching brings the loops to 21 at unchanged accuracy. On code, entropy-matched steering is 0.9 and 1.6 points below CoT@8 at 4B and 8B.

5.3 Controls and the boundary

The mechanism predicts no gain in three situations, evaluated as controls in Table 3. *Multiple choice*: a GPQA-Diamond trace ends in a four-way choice, so closing an open trace yields a guess

Table 4: Trace-state accounting for the paired 4B and 8B cells, first draws, in scope then controls (totals bold, shaded columns ours). Closes and opens are transitions relative to greedy, parentheses give the correct traces. Fixed/broken: answers changed when both close. Term. and Re-dec. exactly decompose steering minus greedy. Fallback: steering fallback wins↔losses.

	n	Greedy	Steering	w↔l	Open	Closes (right)	Opens (right)	Fixed / broken	Term.	Re-dec.	Fallback w↔l
AIME 2024–26	180	126	138	19↔7	35	17 (14)	7 (2)	6 / 5	+11	+1	13↔0
Code	1,400	1,088	1,117	79↔50	153	86 (47)	48 (21)	32 / 29	+26	+3	47↔0
In scope	**1,580**	**1,214**	**1,255**	**98↔57**	**188**	**103 (61)**	**55 (23)**	**38 / 34**	+37	+4	**60↔0**
GPQA-Diamond	396	249	240	30↔39	25	16 (7)	17 (7)	20 / 28	−1	−8	6↔5
Five contest sets	506	267	265	39↔41	103	47 (20)	45 (21)	19 / 17	−4	+2	19↔3
Near ceiling	3,718	3,570	3,551	35↔54	19	13 (12)	27 (24)	27 / 34	−12	−7	8↔0
Controls	**4,620**	**4,086**	**4,056**	**104↔134**	**147**	**76 (39)**	**89 (52)**	**66 / 79**	−17	−13	**33↔8**
All 26 cells	**6,200**	**5,300**	**5,311**	**202↔191**	**335**	**179 (100)**	**144 (75)**	**104 / 113**	+20	−9	**93↔8**

and a re-decision flips among four options. Steering is not distinguishable from greedy: it wins 64 problems and loses 75 over four draws at 4B and 8B ($p = 0.40$) and 13 against 20 at 14B. *Near ceiling*: AMC23, MATH-500 and GSM8K leave under one percent of greedy traces open, so there is nothing to terminate. Steering is half a point below (35 won↔54 lost, $p = 0.06$), the loss from 27 traces it opens. *Beyond competence*: on the five contest sets greedy leaves 15 to 31 percent of traces open, as on AIME, but the derivations are not near completion. Steering is flat (79↔83 over four draws, $p = 0.81$). The fallback's accuracy is not below greedy in any control cell, since it alters only open traces (its eight control losses are greedy traces scored right from unfinished text). Appendix B.4 Table 10 gives per control set results across all scales.

6 Analysis of trace transitions

Trace states. A trace is open if it reaches the token limit and closed if it emits `</think>`. Table 4 pools the 26 paired cells at 4B and 8B (14B and 32B: Tables 6 and 7 in Appendix B) and splits the change in correct answers into termination (traces steering closes, opens or leaves open) and re-decision (answers changed among traces both arms close). Of the 335 traces greedy leaves open, steering is right on 105 and greedy on 20. Steering closes about half of them and opens others, far more closed than opened on AIME and code, about as many on the contest sets and GPQA. Greedy leaves 19 percent of AIME and 20 percent of contest traces open, so neither open rate nor length separates them.

Where each part pays. The termination part is positive only where the closed traces are right: 14 of 17 on AIME (+11) and 47 of 86 on code (+26), against 20 of 47 on the contest sets and 7 of 16 on GPQA (−4 and −1). On the near-ceiling sets steering opens 27 traces, 24 right (−12). The re-decision part is small except on GPQA (−8) and the ceiling sets (−7). Across the 26 cells, accuracy change tracks the change in open traces (Spearman −0.79), not accuracy or open rate (−0.18, +0.30).

Why steering reduces non-termination. A greedy decode maps context to next token deterministically, so a recurring stretch is completed identically each time and the trace cycles to the limit (Holtzman et al., 2020; Fu et al., 2021), while sampling escapes at the torn steps of the cycle. A fight is not a detector of the future loop: steered traces leave the greedy path at their first fight, after a median of 38 tokens at 8B and 77 at 4B (all 180 within 340 tokens), long before any repetition. It marks a branch point where two late-layer decisions disagree (final-layer entropy about 1 nat at applied fights, trough below 0.3). Greedy always takes the final-layer branch, sampling takes another at random, a serving perturbation takes another by numerical accident. Conflict steering selects the alternative deterministically from the model's own intermediate decision. The claim is about populations of paths: in scope, steering closes 103 of the 188 traces greedy leaves open and opens 55, +37 answers through termination (Table 4), and the direction matters, since a second greedy configuration closes 38 of 88 repetition loops at 8B where steering closes 54. The earlier layer leans toward finishing, giving `</think>` more probability than the final layer at fights and proposing step-closing tokens where it hedges (Appendix G). Appendix E.3 Figure 8 shows a representative greedy loop and the corresponding steered trajectory.

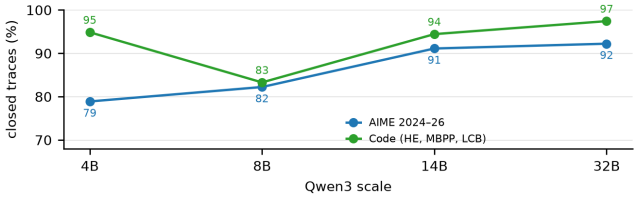

Figure 5: Share of greedy AIME and code traces that close within the token budget, by model (first paired draws). The termination component of the steering gain falls with the open share: $+37$, $+13$ and $+1$ in scope at 4B/8B, 14B and 32B (Table 4, and Tables 6 and 7 in Appendix B).

7 Discussion

When to use it. Figure 5 shows where the regime is: along the 8B, 14B and 32B sweep greedy leaves 16, 8 and 7 of 90 AIME traces open and the termination part of the gain falls from $+5$ to $+3$ and $+1$ (Tables 6 and 7, and Appendix B.1 for the 8B value). The 4B model, with 19 open traces and $+6$ through termination, behaves like 8B. Used as a drop-in replacement for greedy on every task, conflict steering does not change accuracy significantly: over the 26 cells at 4B and 8B it wins 202 problems and loses 191 ($p = 0.61$), the in-scope gains offset by the control losses. *One decode*: where greedy leaves many traces open on generated-answer tasks, steer every decode. At 4B and 8B steering reaches the sampled mean on AIME in three of four serving configurations and improves code. *Steering fallback*: when greedy is the default, rerun with steering only on the traces that reach the token budget without `</think>`. This label-free fallback wins 92 problems and loses none in scope over the four models and recovers as many open traces as one temperature-0.6 sample or a second greedy configuration (see detailed analysis in Appendices C and E.2). *Sampling*: when a vote or verifier and several traces as are affordable. *Cost*: steering decodes the $M - 1$ band layers through the unembedding at every step and our unoptimized fork runs at half of greedy throughput at 4B and 8B (0.79 at 32B, Appendix F), so always-on steering costs about two greedy decodes and Average@8 eight, while the fallback costs about $1 + 2r$ for an open-trace rate r, $1.06\times$ to $1.40\times$ over the observed 3 to 20 percent, spent only where greedy fails to terminate.

Limitations. The AIME gain at 4B and 8B is 3.6 points over four serving configurations and significant only pooled with code, where one steered decode is within one point of the CoT@8 mean at 4B (83.0 against 83.6) and 4.5 points below it at 8B. At 14B and 32B, with under a tenth of the traces open, one steered decode gains 3.3 and 5.6 points on AIME but 0.9 and -1.5 on code (70↔66, $p = 0.80$). At 14B it stays below the sampled mean (72.2 versus 76.1) and the 32B AIME gain comes through re-decisions (Appendix B.1). The fallback gains at every scale. Beyond Qwen3, DeepSeek-R1-Distill-Qwen-14B, a different post-training lineage on a Qwen2.5 base, preserves the band rule and the steering fallback ($+7$ on AIME) but not always-on steering (Appendix B.2), however broader generality remains open. Our reading of the CLD trough as conflict is not causal, and the boundary and the steering fallback were formulated after the first Qwen3 draws.

8 Conclusion

A reasoning model's late layers disagree at rare branch points, and resolving that disagreement in the model's own hidden state selects an alternative deterministic trajectory. Where greedy decoding leaves traces open, those trajectories close far more often. Conflict steering turns this into a deterministic decoding rule with no sampling, verifier or per-model strength. Where greedy leaves traces open it gives one decode the accuracy of a sampled draw on contest math and gains on code. Where nothing is open it changes little, and the steering fallback never loses to greedy in scope at any of the four models for about one extra decode in ten. For anyone who must ship a reproducible trajectory from a reasoning model, that is the practical result: decode greedily, steer the traces that fail to close, and recover most of what sampling would have bought.

AI USE STATEMENT

We used LLM-based assistants to implement parts of the evaluation pipeline and analysis scripts, create or modify figures from our specifications, and draft analyses of stored result files. AI was also used to suggest related work, brainstorm, and edit text for readability. The research question, operator, calibration rule, experimental design, and scientific claims are the authors' own. All AI-assisted code was reviewed and tested by the authors, and reported results were computed from stored result files using reviewed analysis scripts. We manually checked representative outputs and cross-checked aggregate results and reported comparisons for consistency. All cited work was read by the authors, and all AI-assisted edits were reviewed to ensure that they preserved the intended meaning and accurately reflected our methods and results. No assistant was used to generate data or write proofs. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Every number in the paper is computed from stored result files by the analysis scripts of our code base. Section 3 states the operator with all of its constants ($\tau = 0.35$, $k = 14$, $\lambda_{\text{near}} = 1$, $\lambda_{\text{far}} = 0$), Section 4 the rule that fixes the band M from an unlabeled probe, and Appendix A.1 the worked probe. Appendix H lists every method setting per scale (Table 15), the serving configuration of every cell (Table 16), the hardware, software versions and scoring rules. Appendix C defines the paired protocol, the sign test and the bootstrap (Equation 3) and lists every paired draw, because numerical serving variation can change deterministic greedy trajectories and all comparisons are therefore paired within a run. The code, a fork of the Soft Thinking reference implementation on SGLang containing the operator, entropy matching, the calibration probe, will be released as an anonymized repository during the discussion period and publicly with the final version.

REFERENCES

Pranjal Aggarwal, Aman Madaan, Yiming Yang, et al. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12375–12396, 2023.

Georgii Aparin and Tatiana Gaintseva. A geometric account of activation steering through angle–norm decomposition. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026. URL https://openreview.net/forum?id=MVYKmq3Jhx.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-0679. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1d27c01ebd3e3aebe226b44fc970d803-Paper-Datasets_and_Benchmarks_Track.pdf.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

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

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Th6NyL07na.

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

Wilfrid Joseph Dixon and Alexander McFarlane Mood. The statistical sign test. *Journal of the American Statistical Association*, 41 236:557–66, 1946. URL https://api.semanticscholar.org/CorpusID:41692760.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed A Aly, Beidi Chen, and Carole-Jean Wu. Layerskip: Enabling early exit inference and self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12622–12642, 2024. URL https://doi.org/10.18653/v1/2024.acl-long.681.

Hossein Entezari Zarch, Lei Gao, Chaoyi Jiang, and Murali Annavaram. Del: Context-aware dynamic exit layer for efficient self-speculative decoding. In *Proceedings of the Conference on Language Modeling (COLM) 2025*, 2025.

Renyu Fu and Guibo Luo. Selar: Selective latent reasoning in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7073–7087, 2026.

Yichao Fu, Xuewei Wang, Hao Zhang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. In *International Conference on Learning Representations*, volume 2026, pp. 94355–94377, 2026.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. A theoretical analysis of the repetition problem in text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12848–12856, 2021.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15466–15490. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ghandeharioun24a.html.

Raja Gond, Aditya K Kamath, Ramachandran Ramjee, and Ashish Panwar. Llm-42: Enabling determinism in llm inference with verified speculation. In *SOSP 2026*. SOSP, September 2026. URL https://www.microsoft.com/en-us/research/publication/llm-42-enabling-determinism-in-llm-inference-with-verified-speculation/.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Itxz7S4Ip3.

Yunzhen He, Yusuke Takase, and Hidetoshi Shimodaira. Delta: A decoding strategy based on logit trajectory prediction improves factuality and reasoning ability. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pp. 309–321, 2025.

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://openreview.net/forum?id=rygGQyrFvH.

Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, volume 2025, pp. 58791–58831, 2025.

Aayush Karan and Yilun Du. Reasoning with sampling: Your base model is smarter than you think. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.14901.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. ngpt: Normalized transformer with representation learning on the hypersphere. In *International Conference on Learning Representations*, volume 2025, pp. 74014–74038, 2025.

Ming Ma, Bowen Zheng, Zhongqiao Lin, and Tianming Yang. Simlens for early exit in large language models: Eliciting accurate latent predictions with one more token, 2026. URL https://arxiv.org/abs/2507.17618.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, jun 1947. doi: 10.1007/BF02295996.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20286–20332, 2025.

Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs. In *International Conference on Learning Representations*, volume 2025, pp. 70333–70366, 2025.

Georgy Noarov, Soham Mallick, Tao Wang, Sunay Joshi, Yan Sun, Yangxinyu Xie, Mengxin Yu, and Edgar Dobriban. Foundations of top-k decoding for language models. *Advances in Neural Information Processing Systems*, 38:116713–116756, 2026.

nostalgebraist. Interpreting GPT: the logit lens. LessWrong blog post, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Renuka Oladri, Niveda Jawahar, and Abdirisak Mohamed. Token budget saturation and mechanistic early-detection of reasoning non-convergence in chain-of-thought models. In *COLM 2026 Workshop on Efficient Reasoning*, 2026. URL https://openreview.net/forum?id=PpSzvRyl2f.

OpenAI. API reference: Chat completions, `presence_penalty` and `frequency_penalty`. https://platform.openai.com/docs/api-reference/chat/create, 2024. Accessed September 2026.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Taekhyun Park, Yongjae Lee, Dohee Kim, and Hyerim Bae. Loopus: Recasting pretrained llms into looped latent refinement models. *arXiv preprint arXiv:2605.11011*, 2026. URL https://arxiv.org/abs/2605.11011.

Charilaos Pipis, Shivam Garg, Vasilis Kontonis, Vaishnavi Shrivastava, Akshay Krishnamurthy, and Dimitris Papailiopoulos. Wait, wait, wait... why do reasoning models loop? In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=oZWE7mSqlk.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17456–17472, 2022.

Dachuan Shi, Abedelkadir Asi, Keying Li, Xiangchi Yuan, Leyan Pan, Wenke Lee, and Wen Xiao. Swireasoning: Switch-thinking in latent and explicit for pareto-superior reasoning LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=t33kMzEAg8.

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*, volume 2025, pp. 10131–10165, 2025.

Koduvayur Subbalakshmi, Sabbir Hossain Ujjal, Venkata Krishna Teja Mangichetty, and Nastaran Jamalipour Soofi. Listen to the layers: Mitigating hallucinations with inter-layer disagreement. *arXiv preprint arXiv:2602.09486*, 2026.

Yao Tang, Li Dong, Yaru Hao, Qingxiu Dong, Furu Wei, and Jiatao Gu. Multiplex thinking: Reasoning via token-wise branch-and-merge. In *COLM 2026 Workshop on Efficient Reasoning*, 2026. URL https://openreview.net/forum?id=K5IaRy4WmJ.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Rui Wei, Rui Du, Hanfei Yu, Devesh Tiwari, Jian Li, Zhaozhuo Xu, and Hao Wang. The diminishing returns of early-exit decoding in modern llms. *arXiv preprint arXiv:2603.23701*, 2026.

Junhong Wu, Jinliang Lu, Zixuan Ren, Gangqiang Hu, Zhi Wu, Dai Dai, and Hua Wu. LLMs are single-threaded reasoners: Demystifying the working mechanism of soft thinking. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ASLuOoP78o.

Fangzhi Xu, Hang Yan, Chang Ma, Haiteng Zhao, Jun Liu, Qika Lin, and Zhiyong Wu. ϕ-decoding: Adaptive foresight sampling for balanced inference-time exploration and exploitation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13214–13227. Association for Computational Linguistics, July 2025. URL https://aclanthology.org/2025.acl-long.647/.

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/148c0aeea1c5da82f4fa86a09d4190da-Abstract-Conference.html.

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

Zejia You, Chunyuan Deng, and Hanjie Chen. Spherical steering: Geometry-aware activation rotation for language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=9PPeJsVLdB.

Jiayi Yuan, Hao Li, Xinheng Ding, Wenya Xie, Yu-Jhe Li, Wentian Zhao, Kun Wan, Jing Shi, Xia Hu, and Zirui Liu. Understanding and mitigating numerical sources of nondeterminism in llm inference. *Advances in Neural Information Processing Systems*, 38:169819–169851, 2026.

Hongxiang Zhang, Hao Chen, Tianyi Zhang, and Muhao Chen. Active layer-contrastive decoding reduces hallucination in large language model generation. *arXiv preprint arXiv:2505.23657*, 2025a.

Jianyi Zhang, Da-Cheng Juan, Cyrus Rashtchian, Chun-Sung Ferng, Heinrich Jiang, and Yiran Chen. Sled: Self logits evolution decoding for improving factuality in large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/0939f13ffce3ff487509d902ddba4571-Abstract-Conference.html.

Jinyang Zhang, Hongxin Ding, Yue Fang, Weibin Liao, Muyang Ye, Junfeng Zhao, and Yasha Wang. The tell-tale norm: ℓ_2 magnitude as a signal for reasoning dynamics in large language models. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=03ZTlJuX0y.

Jun Zhang, Jue Wang, Huan Li, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. Draft& verify: Lossless large language model acceleration via self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11263–11282, 2024b.

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

Xuanming Zhang, Sining Zhoubian, Yuxuan Chen, Tianyi Tang, An Yang, Sean Du, Chujie Zheng, Fei Huang, Dayiheng Liu, Gao Huang, and Jingren Zhou. Deeper is not always better: Mitigating the alignment tax via confident layer decoding. *arXiv preprint arXiv:2606.21906*, 2026b.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of LLMs in continuous concept space. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=ByQdHPGKgU.

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

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 62557–62583. Curran Associates, Inc., 2024. doi: 10.52202/079017-2000. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/724be4472168f31ba1c9ac630f15dec8-Paper-Conference.pdf.

A CALIBRATION AND ABLATIONS

A.1 OPERATOR AND LAYER PROBE

Table 5 and Figure 6 give the late-layer probe. The band rule was developed from sweeps at the two calibration scales. At 8B, the earlier trough-mixing operator (Soft CLD) over $M \in \{3, 4, 5\}$ gave 56, 61 and 60 of 90 on AIME-3yr: accuracy is highest when the band starts at the first layer with $\overline{\mathrm{JSDn}} < 0.1$ (layer 32, $M = 4$), lower when that layer is excluded ($M = 3$ keeps only the near-copies of the final layer, which detect few conflicts) and when the still-distant layer 31 is added ($M = 5$). The rule was then applied unchanged to Qwen3-14B and Qwen3-32B. At every scale, normalized JSD to the final layer drops sharply near the output. The band starts at the first layer below 0.10 (Section 4). Figure 7 shows the disagreement d at fights. Its distribution is similar across contest math, science and code: the 8B median is 0.433 on AIME 2026 and 0.443 on GPQA, and 28 percent of fights fall below 0.35 on AIME, the five contest sets and code. The overlaid $\alpha(d)$ shows that τ controls the interpolation strength rather than separating two classes of fights.

Table 5: Layer probe on AIME 2024 (30 problems at 8B and 14B, 10 at 4B and 32B). JSDn is the normalized JSD to the final layer, taken as the median over the probed decode steps of each problem and averaged over problems. Agree is the mean top-1 agreement with the final layer. Bold marks the selected band.

Model	quantity	layer (final F on the right)							
4B, $F = 35$	layer	28	29	30	**31**	**32**	**33**	**34**	**35**
	JSDn	0.53	0.21	0.13	0.062	0.021	0.003	0.001	0.000
	agree	0.45	0.57	0.62	0.68	0.76	0.85	0.91	1.00
8B, $F = 35$	layer	28	29	30	31	**32**	**33**	**34**	**35**
	JSDn	0.83	0.55	0.38	0.21	0.093	0.010	0.002	0.000
	agree	0.33	0.45	0.51	0.58	0.65	0.78	0.88	1.00
14B, $F = 39$	layer	33	34	35	**36**	**37**	**38**	**39**	
	JSDn	0.52	0.30	0.18	0.058	0.014	0.003	0.000	
	agree	0.45	0.54	0.60	0.70	0.80	0.90	1.00	
32B, $F = 63$	layer	58	59	60	**61**	**62**	**63**		
	JSDn	0.85	0.69	0.33	0.083	0.004	0.000		
	agree	0.33	0.42	0.58	0.73	0.89	1.00		

Reading M from the probe. At 4B (30 AIME 2024 prompts decoded with the Soft Thinking protocol, and a greedy probe gives the same band), layer 29 is the last point on the sharp distance drop, with $\overline{\mathrm{JSDn}} = 0.21$, and is excluded. Layer 31 is the first point after the curve flattens, with $\overline{\mathrm{JSDn}} = 0.062$ and top-1 agreement 0.681, so the band is layers 31–35 and $M = 5$ (Figure 6). The disagreement threshold τ does not enter this choice.

Threshold τ . We test $\tau = 0.35$ and against $\tau \in \{0.25, 0.45\}$ at 4B to verify it is the optimum configuration. Paired against the default cell on the same problems and batch setting: $\tau = 0.25$ (mean applied weight 0.25) gives 76.67% (69/90) on AIME-3yr against 81.11% (73/90), and 64.14% (127/198) against 64.65% (128/198) on GPQA. A threshold $\tau = 0.45$ gives 77.78% (70/90) on AIME-3yr, and 65.15% (129/198).

B FULL DETERMINISTIC RESULTS

B.1 QWEN3-14B AND QWEN3-32B

Qwen3-14B (40 blocks, tensor parallel 2) is calibrated by the same rule to $M = 4$ (band 36–39, Appendix A.1). Greedy leaves 8 of 90 AIME-3yr traces open, 39 of 700 code traces and 7 of 198 on GPQA, where steering loses slightly (13 won↔20 lost), as at the other scales. Table 6 gives the first paired draws: steering gains 3 on AIME (8↔5) and 6 on code (32↔26), all from terminations ($+13$ in scope) against a small re-decision loss (-4). The steering fallback wins 24 and loses none.

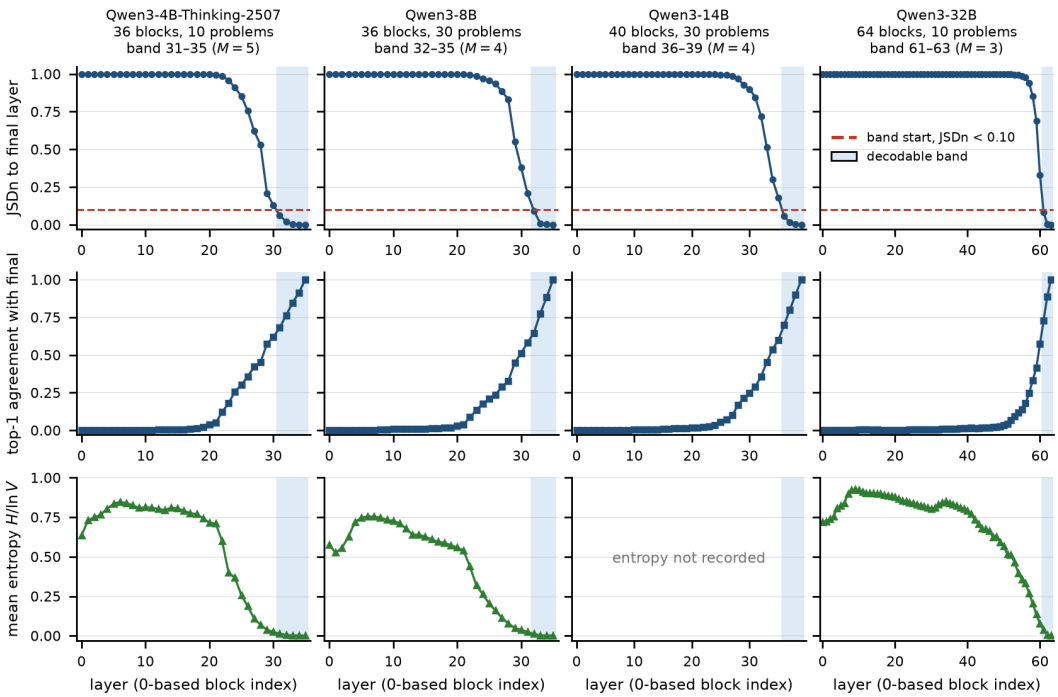

Figure 6: Layer probe at full depth on AIME 2024 (30 problems at 8B and 14B, 10 at 4B and 32B) for the four Qwen3 models. Top: normalized JSD of each layer's next-token distribution to the final layer's, with the band-start line at 0.10 and the selected band shaded (the last layers are Figure 4). Middle: mean top-1 agreement with the final layer. Bottom: mean next-token entropy over the vocabulary, $H/\ln V$ (not recorded at 14B). Entropy collapses several layers before the distributions agree, which is why the band is set from the divergence rather than from the entropy.

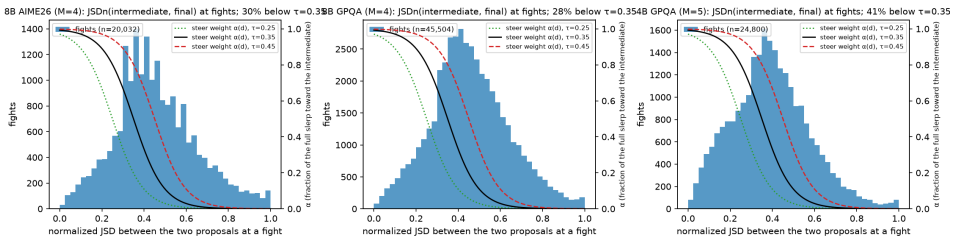

Figure 7: Normalized JSD between the intermediate and final distributions at fights, with the steering weight $\alpha(d)$ overlaid for $\tau \in \{0.25, 0.35, 0.45\}$.

Qwen3-32B (64 blocks) is calibrated by the same rule to $M = 3$ (band 61–63, Appendix A.1). Its greedy decodes rarely fail to terminate: 7 of 90 AIME-3yr traces are open (5 of them loops), 13 of 279 on LiveCodeBench, 2 of 164 on HumanEval and 3 of 257 on MBPP, against 21 of 198 on GPQA. The termination mechanism therefore has little to act on, and Table 7 shows ramaining trace statistics. On AIME steering still gains, 72 against 67 (10↔5) and 68 against 66 in a second draw at batch 12 (7↔5), but through re-decisions (+4) rather than termination (+1). One steered decode scores 80.0 percent against a CoT@8 mean of 76.7 (Table 12). GPQA is unchanged within noise (17 won↔20 lost, 62.6 against 64.1, both below the sampled mean of 67.0). HumanEval and MBPP remain near ceiling. LiveCodeBench loses (11↔22 in the first draw, 30↔43 pooled over two draws, $p = 0.16$). The presence-penalty arms are 67 (greedy), 68 ($\beta = 1.0$), 72 (steering) and 69 (steering with $\beta = 1.0$) of 90. Steering costs 5 percent of the forward pass at 32B and runs at 0.79 of greedy throughput. One decode gains little in this regime. The steering fallback remains non-negative here (26↔4 over the 13 cells).

Table 6: Qwen3-14B, first draws of the in-scope cells and GPQA-Diamond, in the format of Table 4.

	n	Greedy	Steering	w↔l	Open	Closes (right)	Opens (right)	Fixed / broken	Term.	Re-dec.	Fallback w↔l
AIME 2024–26	90	62	65	8↔5	8	5 (5)	4 (1)	4 / 4	+3	0	4↔0
Code	700	569	575	32↔26	39	26 (21)	26 (10)	12 / 16	+10	−4	20↔0
GPQA-Diamond	198	119	112	13↔20	7	7 (2)	5 (4)	11 / 17	−1	−6	2↔1
In scope	790	631	640	40↔31	47	31 (26)	30 (11)	16 / 20	+13	−4	24↔0

Table 7: Qwen3-32B, first draws of the 13 paired cells, in the format of Table 4.

	n	Greedy	Steering	w↔l	Open	Closes (right)	Opens (right)	Fixed / broken	Term.	Re-dec.	Fallback w↔l
AIME 2024–26	90	67	72	10↔5	7	4 (4)	6 (3)	6 / 2	+1	+4	4↔0
Code	700	594	584	20↔30	18	17 (4)	13 (4)	16 / 26	0	−10	4↔0
GPQA-Diamond	198	127	124	17↔20	21	14 (6)	15 (6)	11 / 11	−3	0	5↔4
Five contest sets	253	141	141	22↔22	21	13 (5)	21 (8)	16 / 14	−2	+2	6↔0
Near ceiling	1,859	1,787	1,796	27↔18	8	8 (8)	10 (10)	20 / 10	−1	+10	7↔0
All	3,100	2,716	2,717	96↔95	75	56 (27)	65 (31)	69 / 63	−5	+6	26↔4

B.2 Transfer to a distilled reasoning model

DeepSeek-R1-Distill-Qwen-14B (Guo et al., 2025) is a reasoning model from a different post-training lineage, distilled from DeepSeek-R1 onto a Qwen2.5-14B base (48 blocks). We applied the procedure of Sections 3 and 4 unchanged. The unlabeled probe (30 stored greedy AIME 2024 traces, 7,650 positions) gives $\overline{\mathrm{JSDn}}$ of 0.215, 0.068, 0.013 and 0.003 for layers 43 to 46, so the band rule selects $M = 4$ (layers 44 to 47), a compact band like 32B. The threshold τ, k and the doses are those of Section 3. Greedy decoding at batch 16 (tensor parallel 2) leaves 34 of 90 AIME traces open, 32 of them repetition loops, the regime in which steering gains at 4B and 8B. The account's prediction, recorded before the steered run, was a gain of at least two AIME problems and a clearly positive fallback.

Table 8 gives the outcome. The termination half of the account transfers: steering brings the open AIME traces from 34 to 23 and is right on 10 of the 34 greedy-open problems against 3 for greedy. The re-decision half does not: among the 45 AIME problems, steering's net losses are 8, and on GPQA-Diamond it opens more traces than it closes (29 to 37) and with net losses of 16 among the closed ones. One always-on steered decode therefore is below greedy on AIME within one problem (47 versus 48, 10↔11) and 16 problems down on GPQA-Diamond (97 versus 113, 12↔28, $p = 0.017$). The steering fallback transfers: 55 of 90 on AIME (61.1 percent, +7 over greedy for 1.38 decodes per problem), above six of the eight CoT@8 runs (mean 58.8, standard deviation 3.8, worst 55.6, all judged by the same rule), and equal to greedy on GPQA-Diamond (113 each). On this model conflict steering is not a drop-in for greedy, but the fallback is.

Table 8: DeepSeek-R1-Distill-Qwen-14B at batch 16 with $M = 4$ from the probe: correct answers (traces open at the token limit) for one greedy and one steered decode, problems won↔lost by steering, and the steering fallback. Shaded: ours. Bold: best per row, ties both bold.

Set	n	Greedy (open)	Conflict steering (open)	w↔l	Steering fallback
AIME 2024	30	18 (10)	19 (5)	4↔3	**21**
AIME 2025	30	**18** (10)	15 (8)	2↔5	**18**
AIME 2026	30	12 (14)	13 (10)	4↔3	**16**
AIME-3yr	90	48 (34)	47 (23)	10↔11	**55**
GPQA-Diamond	198	**113** (29)	97 (37)	12↔28	**113**

B.3 Code

Table 9 gives the code results per set for the four models. A solution is the fenced Python block that defines it (for HumanEval, the block containing the entry point), executed against the benchmark's tests.

At 4B, all three code sets improve (5↔2, 7↔2 and 20↔14). At 8B, HumanEval gains two problems (5↔3), MBPP gains seven (13↔6), and LiveCodeBench gains six (29↔23), five of them on its hard

Table 9: Code in full: greedy decoding at batch size 16 (14B: 32, and 24 on LiveCodeBench) and Average@8 $\pm$ the standard deviation of the per-run accuracy at temperature 0.6 (CoT@8 at 4B, 8B and 14B; steering with top-20 entropy matching at 4B and 8B; Soft Thinking was not run on code, ST+Gumbel only on HumanEval and MBPP). Cells report accuracy in percent (correct/n, no-</think> percent). Steering cells add problems won↔lost against greedy, with the two-sided sign test on the pooled column. Bold: best per column and scale, sampled rows included (ties both bold). Underlined: best deterministic result where a sampled mean is higher. Steering fallback rows are the two-decode fallback.

		HumanEval	MBPP	LiveCodeBench v5	Pooled
4B	Greedy	95.1 (156/164, 1.2)	92.6 (238/257, 0.8)	62.0 (173/279, 11.5)	81.0 (567/700, 5.1)
	Conflict steering	**97.0** (159/164, 0.6); 5↔2	94.6 (243/257, 0.0); 7↔2	64.2 (179/279, 11.1); 20↔14	83.0 (581/700, 4.6); 32↔18 ($p = 0.07$)
	CoT @8	96.3 $\pm$ 1.2 (1264/1312, 0.5)	**94.9** $\pm$ 0.8 (1951/2056, 0.1)	**65.8** $\pm$ 1.1 (1469/2232, 7.5)	**83.6** $\pm$ 0.6 (4684/5600, 3.1)
	Soft Thinking @8	—	—	—	—
	ST+Gumbel @8	96.0 $\pm$ 1.3 (1260/1312, 0.1)	94.1 $\pm$ 0.9 (1934/2056, 0.0)	—	—
	Conflict steering, Em @8	96.1 $\pm$ 0.6 (1261/1312, 0.3)	94.1 $\pm$ 1.1 (1934/2056, 0.1)	64.4 $\pm$ 1.5 (1437/2232, 8.1)	82.7 $\pm$ 0.6 (4632/5600, 3.4)
8B	Greedy	87.2 (143/164, 9.1)	89.1 (229/257, 9.3)	53.4 (149/279, 28.0)	74.4 (521/700, 16.7)
	Conflict steering	88.4 (145/164, 6.7); 5↔3	91.8 (236/257, 6.2); 13↔6	55.6 (155/279, 20.1); 29↔23	76.6 (536/700, 11.9); 47↔32 ($p = 0.11$)
	CoT @8	**93.4** $\pm$ 1.1 (1226/1312, 1.1)	**95.8** $\pm$ 0.1 (1969/2056, 1.0)	**60.2** $\pm$ 1.6 (1344/2232, 3.0)	**81.1** $\pm$ 0.6 (4539/5600, 1.8)
	Soft Thinking @8	—	—	—	—
	ST+Gumbel @8	92.6 $\pm$ 1.5 (1215/1312, 0.0)	95.1 $\pm$ 0.5 (1955/2056, 0.0)	—	—
	Conflict steering, Em @8	90.7 $\pm$ 2.0 (1190/1312, 4.5)	94.3 $\pm$ 0.9 (1939/2056, 2.9)	59.1 $\pm$ 1.3 (1319/2232, 9.4)	79.4 $\pm$ 0.7 (4448/5600, 5.8)
14B	Greedy	94.5 (155/164, 4.3)	94.2 (242/257, 2.7)	61.6 (172/279, 9.0)	81.3 (569/700, 5.6)
	Conflict steering	95.1 (156/164, 3.0); 6↔5	94.6 (243/257, 1.9); 6↔5	63.1 (176/279, 10.4); 20↔16	82.1 (575/700, 5.6); 32↔26 ($p = 0.51$)
	Steering fallback	**98.2** (161/164, 0.0)	**96.1** (247/257, 0.8)	**64.9** (181/279, 3.9)	**84.1** (589/700, 1.9)
	CoT @8	96.6 $\pm$ 1.8 (1268/1312, 0.7)	95.9 $\pm$ 0.9 (1972/2056, 0.2)	64.6 $\pm$ 1.1 (1442/2232, 1.0)	83.6 $\pm$ 0.9 (4682/5600, 0.6)
32B	Greedy	95.7 (157/164, 1.2)	97.3 (250/257, 1.2)	67.0 (187/279, 4.7)	84.9 (594/700, 2.6)
	Conflict steering	**96.3** (158/164, 0.6); 5↔4	97.3 (250/257, 0.4); 4↔4	63.1 (176/279, 4.3); 11↔22	83.4 (584/700, 2.0); 20↔30 ($p = 0.20$)
	Steering fallback	**96.3** (158/164, 0.6)	**98.1** (252/257, 0.0)	**67.4** (188/279, 0.0)	**85.4** (598/700, 0.1)

tier. The 8B traces are also shorter and fewer remain open. Against sampling, one steered decode is within one point of the CoT@8 mean at 4B (83.0 against 83.6 pooled) and 4.5 points below it at 8B (76.6 against 81.1): sampling closes most of the 8B traces that greedy leaves open (3.0 percent of samples truncated against 16.7 percent of greedy decodes), which steering only partly does. Steering under the sampler with entropy matching stays below CoT@8 on HumanEval and MBPP at both scales, and at 8B it truncates 59 HumanEval samples (40 loops) against 15 for CoT@8.At 32B, HumanEval and MBPP are near ceiling and change by at most one problem, and LiveCodeBench loses 30↔43 over two draws ($p = 0.16$), concentrated on the hard split. Code fight rates range from 0.037 to 0.072 per token, with the same share of near fights as AIME.

B.4 Controls per benchmark and other operators

Table 10: Controls per benchmark: accuracy in percent for one greedy decode, one conflict-steered decode and the steering fallback at the paired batch size (first draws), and the CoT@8 mean $\pm$ standard deviation where sampling was run (GPQA-Diamond at 4B, 8B and 32B). At 14B only GPQA-Diamond was run. Pooled values are in Table 3. Shaded: ours. Bold: best per column and scale, ties both bold. Underlined: best deterministic result where a sampled mean is higher.

		Multiple choice	Beyond competence					Near ceiling		
Model	Method	GPQA-Diamond (198)	HMMT (30)	BRUMO (30)	CMIMC (40)	SMT (53)	BeyondAIME (100)	AMC23 (40)	MATH-500 (500)	GSM8K (1,319)
4B	Greedy	**66.7**	53.3	66.7	65.0	71.7	44.0	**100.0**	**98.2**	**95.7**
	CoT @8	65.5 $\pm$ 1.7	—	—	—	—	—	—	—	—
	Conflict steering	64.6	46.7	**80.0**	47.5	**75.5**	**46.0**	**100.0**	97.8	95.1
	Steering fallback	**66.7**	**56.7**	73.3	**67.5**	73.6	**46.0**	**100.0**	**98.2**	**95.7**
8B	Greedy	59.1	36.7	**73.3**	45.0	64.2	38.0	92.5	97.0	95.1
	CoT @8	59.1 $\pm$ 2.3	—	—	—	—	—	—	—	—
	Conflict steering	56.6	**50.0**	66.7	37.5	58.5	41.0	87.5	96.4	94.8
	Steering fallback	**59.6**	43.3	**73.3**	**47.5**	**67.9**	**42.0**	**97.5**	**97.6**	**95.4**
14B	Greedy	60.1	—	—	—	—	—	—	—	—
	Conflict steering	56.6	—	—	—	—	—	—	—	—
	Steering fallback	**60.6**	—	—	—	—	—	—	—	—
32B	Greedy	64.1	63.3	80.0	50.0	62.3	**45.0**	**97.5**	96.6	95.9
	CoT @8	**67.0** $\pm$ 2.3	—	—	—	—	—	—	—	—
	Conflict steering	62.6	60.0	**83.3**	50.0	62.3	**45.0**	**97.5**	97.2	**96.4**
	Steering fallback	64.6	**70.0**	80.0	**55.0**	**66.0**	**45.0**	**97.5**	**97.4**	96.1

Table 10 gives every control benchmark at every scale for greedy, one steered decode and the steering fallback, with the CoT@8 mean on GPQA-Diamond. AMC23, MATH-500 and GSM8K, on which

greedy already scores 92 to 100 percent, are within two points of their ceiling for every method, and conflict steering changes them by at most one point (paired wins to losses against greedy over three benches and three scales 62↔72, $p = 0.44$).

C PAIRED STATISTICS AND SERVING SENSITIVITY

Across every greedy-path decode we ran, the greedy points alone span five problems at 8B, and no consistent relation between open traces and correct answers survives across draws.

A second greedy configuration as the fallback. A greedy path changes only when the serving configuration changes, so the closest analogue of "rerun with another seed" is the second greedy draw of Table 11. On the 18 cells with two greedy draws at 4B and 8B (AIME, GPQA-Diamond and the five contest sets, 1,082 problems), the second draw closes 77 of the 163 traces the first leaves open and is right on 39 of them, steering closes 80 and is right on 41, and one temperature-0.6 sample closes 46 and is right on 25 of the 60 AIME and GPQA traces in expectation, against 33 and 24 for steering. On the open traces alone, then, any change of path recovers about the same number: on AIME 16 of 35 for the second configuration, 16 for one sample and 15 for steering. The two kinds of open trace differ, however. All 83 open 4B traces wander without repeating, and there the second configuration and steering close the same number (32 and 31). At 8B, 88 of the 119 open traces are repetition loops (with the two 8B code cells that have a second greedy draw included): the second configuration closes 38 of them, since a change of batch composition alters the path only through numerical noise and the model re-enters the same cycle, while steering closes 54 (on HumanEval and MBPP, 19 against 6 of 36). On the 31 wandering 8B traces the two close 13 and 16. The difference lies in what happens to the traces the first draw closed: the second greedy draw opens 74 of them, 40 previously right, so as a drop-in it reshuffles (accuracy 642 to 634 over the 1,082 problems). Steering opens 69, 30 previously right (642 to 643 over the same mixed set, which is dominated by the controls), and on the in-scope tasks closes far more than it opens (17 against 7 on AIME, 86 against 48 on code, Table 4). Steering is therefore not immune to non-termination. It is a deterministic tie-break that closes wandering traces as often as any change of path, closes repetition loops more often, and opens fewer traces than a random change of path does, in the same decode.

Second draws of the five contest sets. Repeating the contest cells of Table 3 at a second batch size at 4B and 8B, per-set differences change sign between draws (BeyondAIME at 4B $+2$ then -6, HMMT at 8B $+4$ then -5) while the pooled result stays flat: 101↔105 over 1,265 paired problems in five draws ($p = 0.83$). This is the noise floor of Section 5.1 seen on harder sets, and the reason the paper reports these sets pooled rather than cell by cell.

Table 4 in the main text pools the 26 cells at 4B and 8B and Table 7 the 13 cells at 32B. Over all 39 cells the drop-in comparison is 298↔286 (412↔403 with the second draws, 10,751 pairs) and the steering fallback wins 119 problems and loses 12 (a gain of 147 problems with the second draws). On the open AIME problems the steered rerun is right on 8 of 19, 7 of 16 and 4 of 7 at 4B, 8B and 32B, against a mean of 7.1, 8.9 and 3.2 for one temperature-0.6 sample. On GPQA 9 of 25 and 6 of 21 at 8B and 32B against 8.6 and 7.1. The fallback's AIME-3yr accuracy is 83.3, 71.1 and 78.9 percent.

Statistics. Every paired comparison uses a two-sided exact sign test on discordant problems, the exact form of McNemar's test for paired binary outcomes (McNemar, 1947; Dixon & Mood, 1946). Let w be the number of problems solved only by steering and l the number solved only by the baseline. We omit ties, where both methods are right or both are wrong. Under the null hypothesis that a changed trace is equally likely to fix or break an answer, $w \sim \mathrm{Bin}(w + l, \frac{1}{2})$ and

$$p \;=\; \min\Big(1,\; 2^{\,1-(w+l)} \sum_{i=0}^{\min(w,l)} \binom{w+l}{i}\Big). \tag{3}$$

The 4B batch-20 draw is the selection draw on which the operator and its constants were chosen, so it is not an independent confirmation. Without it the AIME result is 25↔17 ($p = 0.28$). For the pooled AIME draws in Table 11, $w = 33$ and $l = 20$ give $p = 0.10$, and with the two code cells, $w = 112$ and $l = 70$ give $p = 0.002$. The effect size is the net difference over all pairs,

Table 11: Paired steering and greedy draws at the same batch size and code state on the in-scope tasks. Wins and losses count problems solved by exactly one method. The value p is the two-sided sign test, the chance of a split at least this uneven if steering had no effect.

Cell	Batch	Steering vs greedy (correct)	wins↔losses	p
4B AIME-3yr	20	73 vs 68	8↔3	0.23
4B AIME-3yr	16	71 vs 68	7↔4	0.55
8B AIME-3yr	12	65 vs 58	11↔4	0.12
8B AIME-3yr	16	61 vs 63	7↔9	0.80
AIME pooled, 360 pairs		270 vs 257	33↔20	0.10
4B code (700)	16	581 vs 567	32↔18	0.07
8B code (700)	16	536 vs 521	47↔32	0.11
4B and 8B in scope, 1,760 pairs		**1,387 vs 1,345**	**112↔70**	**0.002**
14B AIME-3yr	30	65 vs 62	8↔5	0.58
14B code (700)	32, 24	575 vs 569	32↔26	0.51
32B AIME-3yr	16	72 vs 67	10↔5	0.30
32B AIME-3yr	12	68 vs 66	7↔5	0.77
32B code (700)	16	584 vs 594	20↔30	0.20
Four scales in scope, 3,430 pairs		**2,751 vs 2,703**	**189↔141**	**0.010**

$(w-l)/360 = 0.036$ on AIME. The four AIME draws reuse the same 90 problems, so the 360 pairs are not independent. As a conservative check, we sum outcomes for each (model, problem) pair and test the sign of its net result: 23 pairs favor steering and 13 favor greedy out of 180 ($p = 0.13$). The corresponding GPQA result is 42 against 52 out of 396 ($p = 0.35$). With the two 32B draws of Appendix B.1 the six-draw AIME result is 50↔30 over 540 pairs ($p = 0.033$, 37 against 20 of 270 units).

Intervals for Average@8 differences are percentile bootstrap intervals over problems (Efron & Tibshirani, 1993). We resample problems, compute the paired difference of their mean accuracies, and take the 2.5th and 97.5th percentiles. The ridges in Figure 1 instead draw one of eight samples for each problem to estimate the accuracy distribution of one sampled decode. Following recommendations for paired NLP evaluation (Dror et al., 2018), we describe results below $p = 0.05$ as directional findings, report counts without a general claim between 0.05 and 0.15, and call larger values flat.

D Sampling and entropy matching

D.1 Sampled diagnostics

Table 12 gives the sampled results in full at every scale (pass@8, majority, truncated samples and loops). We label a truncated sample as a repetition loop when at least half of the 8-grams in its last 3,000 words are repeats (8-gram repetition ratio at least 0.5). A wandering trace has a ratio below 0.2. On 8B GPQA, steering produces 42 truncated samples out of 1,584, including 39 loops, compared with 5 truncations and 4 loops for chain-of-thought. Accuracy remains similar: Average@8 is 58.02 versus 59.09, pass@8 is 81.3 versus 80.3, and majority accuracy is 57.6 versus 57.1. At 4B, full-vocabulary entropy matching gives 82.92, 72.50 and 75.42 on the three AIME years, within four samples of plain steering. None at 4B scale produces a detected repetition loop. At 32B, top-30 entropy matching on GPQA-Diamond gives 65.78 against 66.98 for chain-of-thought (paired difference -1.2 points, 38 problems above, 53 below, bootstrap interval -3.2 to $+0.7$), with 73 truncated samples (67 loops) against 3 (none): the loop excess is largest at the scale whose sampler uses the widest support.

D.2 Entropy matching on the sampled support

The trough's stop signal. Re-prefilling the first 4,096 generated tokens of the 90 paired greedy AIME traces per scale on CPU and decoding the band at every step: the trough layer assigns more probability to `</think>` than the final layer at 23 to 25 percent of trough steps at 4B and 31 to 37 percent at 8B. At fights the mean of $\log p_I(\texttt{</think>}) - \log p_F(\texttt{</think>})$ is $+0.86$ at 4B and $+1.5$ at 8B ($+1.4$ and $+2.0$ at near fights) with a median near zero, so at most fights both layers give the stop token negligible mass and, when they differ, it is the trough that favors stopping. Neither

Table 12: Sampled decoding in full, AIME-3yr (720 samples) and GPQA-Diamond (1,584): Average@8 $\pm$ the standard deviation of the per-run accuracy (correct samples, no-</think> percent), pass@8, maj@8 (majority vote over the eight extracted answers, ties to the earliest), and truncated samples (loops). Em matches final-layer entropy at each fight on the sampler's top-k support. The full-vocabulary rows match the whole distribution (Appendix D.2). @1 rows repeat the deterministic decodes of Table 1. Bold: best per column and scale. Underlined: best single decode where a sampled mean is higher.

Model	Method	AIME-3yr	pass@8	maj@8	truncated (loops)	GPQA
4B	Greedy @1	75.56 (68, 21.1)	—	—	19 (0)	**66.67**
	Conflict steering @1	**81.11** (73, 13.3)	—	—	12 (0)	64.65
	CoT @8	77.22 $\pm$ 3.2 (556, 15.0)	90.0	**86.7**	108 (0)	65.47 $\pm$ 1.7
	Soft Thinking @8	74.31 $\pm$ 3.3 (535, 12.6)	**91.1**	84.4	91 (0)	64.08 $\pm$ 1.7
	ST+Gumbel @8	76.39 $\pm$ 3.2 (550, 11.7)	87.8	85.6	84 (0)	63.57 $\pm$ 1.4
	Conflict steering @8	77.92 $\pm$ 2.3 (561, 15.3)	88.9	**86.7**	110 (0)	64.90 $\pm$ 2.4
	Conflict steering, Em @8	76.94 $\pm$ 1.9 (554, 15.4)	90.0	85.6	111 (0)	—
	Conflict steering, Em (full vocabulary) @8	76.94 $\pm$ 2.9 (554, 15.3)	86.7	85.6	110 (0)	65.21 $\pm$ 1.2
8B	Greedy @1	64.44 (58, 17.8)	—	—	16 (11)	59.09
	Conflict steering @1	**72.22** (65, 14.4)	—	—	13 (5)	56.57
	CoT @8	69.86 $\pm$ 3.0 (503, 7.5)	84.4	76.7	54 (4)	59.09 $\pm$ 2.3
	Soft Thinking @8	64.31 $\pm$ 1.9 (463, 5.6)	77.8	74.4	40 (0)	56.44 $\pm$ 2.2
	ST+Gumbel @8	69.72 $\pm$ 2.7 (502, 6.1)	**87.8**	**78.9**	44 (0)	**60.42** $\pm$ 1.9
	Conflict steering @8	68.19 $\pm$ 2.5 (491, 13.6)	83.3	76.7	98 (35)	58.02 $\pm$ 1.9
	Conflict steering, Em @8	68.75 $\pm$ 3.4 (495, 10.3)	83.3	77.8	74 (21)	58.08 $\pm$ 2.6
	Conflict steering, Em (full vocabulary) @8	69.03 $\pm$ 2.5 (497, 11.7)	83.3	77.8	84 (26)	—
14B	Greedy @1	68.89 (62, 8.9)	—	—	8 (5)	—
	Conflict steering @1	72.22 (65, 7.8)	—	—	7 (3)	—
	CoT @8	**76.11** $\pm$ 2.7 (548, 5.0)	90.0	**85.6**	36 (5)	—
32B	Greedy @1	74.44 (67, 7.8)	—	—	7 (5)	64.14
	Conflict steering @1	**80.00** (72, 10.0)	—	—	9 (4)	62.63
	CoT @8	76.67 $\pm$ 3.1 (552, 6.2)	**91.1**	**86.7**	45 (9)	**66.98** $\pm$ 2.3
	Soft Thinking @8	74.58 $\pm$ 2.2 (537, 3.6)	85.6	81.1	26 (0)	64.46 $\pm$ 2.1
	ST+Gumbel @8	74.86 $\pm$ 3.5 (539, 5.1)	88.9	82.2	37 (0)	64.08 $\pm$ 1.2
	Conflict steering, Em @8	76.39 $\pm$ 2.0 (550, 6.5)	87.8	84.4	47 (13)	65.78 $\pm$ 3.5

layer's top token is </think> within the first 4,096 tokens. This is a tendency at conflict steps, not the closing event itself.

The matching step as implemented. At each applied fight the temperature solves

$$H\big(\mathrm{softmax}(\boldsymbol{z}'/t)\big) = H\big(p^{(F)}\big), \tag{4}$$

by bisection on $\ln t$ (24 steps), since entropy increases monotonically with t. Matching on the full vocabulary instead of the sampler's top-k support changes accuracy by at most two problems of 720 on AIME at 4B and 8B (Table 12). The top-k variant matches the distribution the sampler draws from and leaves fewer loops (21 against 26 at 8B, Figure 3). On every row of the batch at once, t is searched in range $[0.02, 50]$, and a target outside the reachable range ends at the nearest bound. The full-vocabulary variant (rows marked "full vocabulary" in Table 12) sets $K = V$. Under greedy decoding a positive temperature leaves the argmax unchanged, so the step is skipped and steered greedy traces are identical with or without matching. The mean t over applied 8B fights is 1.40 on the top-20 support.

Repetition under sampling. The trough has the lowest entropy in the band, so interpolation sharpens the next-token distribution at most fights: across 592,652 fights in the sampled 8B AIME runs the final layer averages 0.99 nats, the trough is below 0.3, and the steered distribution is sharper than the final layer at 70 percent of fights, by 0.23 nats on average (Figure 3). Greedy decoding is unaffected beyond the argmax, but under sampling reduced entropy makes repeated states harder to escape. Loops fall with how much of the sampler-visible entropy is restored: 35 (raw steer), 26 (full-vocabulary match) and 21 (top-20 match) against 4 for chain-of-thought over 720 samples, truncations 98, 84 and 74 against 54, at unchanged accuracy (68.19, 69.03 and 68.75 against 69.86). At 4B no arm loops and matching changes nothing (76.94 against 77.92). On GPQA-Diamond at 8B the top-20 match brings loops from 39 to 33 against 4 over 1,584 samples at unchanged accuracy

(58.02 and 58.08 against 59.09), and at 4B full-vocabulary matching gives 65.21 against 64.90 raw and 65.47 for chain-of-thought.

E ALTERNATIVE REPAIRS AND MECHANISM CONTROLS

E.1 PRESENCE-PENALTY ARMS IN FULL

Results. A presence penalty subtracts a fixed β from the logit of every token generated so far, once, however often the token recurs (the additive penalty of the OpenAI API (OpenAI, 2024) as implemented in SGLang). We test $\beta \in \{0.5, 1.0\}$, inside the range the Qwen3 model card recommends (Yang et al., 2025), on both arms (Table 13). The effect depends on the model: 1.0 helps at 8B (58 to 63) and hurts at 4B (68 to 66), where 0.5 helps instead (74). Choosing the better strength per scale on the test labels gives 137 correct of 180, either strength fixed across scales gives 129, and steering gives 138 with its band from the unlabeled probe (Section 4). Paired on AIME-3yr, steering without a penalty beats greedy with $\beta = 1.0$ by 25↔12 pooled over the three scales ($p = 0.047$). On 8B code, $\beta = 1.0$ gives 544 of 700 against 521 for greedy (53↔30, $p = 0.015$) and 536 for conflict steering (36↔28, $p = 0.38$), and the steering fallback gives 559 (37↔22 against the penalty, $p = 0.07$). A penalty added to steering lowers accuracy at both scales, GPQA stays within noise, and under sampling $\beta = 1.0$ moves 8B chain-of-thought from 69.9 to 71.5, since sampling already closes traces.

Table 13: All presence-penalty arms, greedy decoding: AIME-3yr correct of 90 with open traces (loops) and GPQA-Diamond correct of 198 with open traces. The 8B code column pools HumanEval, MBPP and LiveCodeBench (700 problems, batch 16): correct (open). Shaded: ours. Bold: best per column, ties both bold.

	4B		8B			32B	
Arm	AIME-3yr	GPQA	AIME-3yr	GPQA	Code (700)	AIME-3yr	GPQA
Greedy	68 (19, 0)	**132** (0)	58 (16, 11)	**117** (25)	521 (117)	67 (7, 5)	**127** (21)
Greedy, pp 0.5	**74** (13, 0)	130 (0)	55 (7, 4)	114 (16)	—	—	—
Greedy, pp 1.0	66 (17, 0)	**132** (1)	63 (4, 2)	112 (10)	**544** (58)	68 (8, 2)	**127** (4)
Conflict steering	73 (12, 0)	128 (0)	**65** (13, 5)	112 (26)	536 (83)	**72** (9, 4)	124 (22)
Steering, pp 0.5	66 (18, 0)	129 (0)	60 (15, 8)	113 (26)	—	—	—
Steering, pp 1.0	68 (15, 0)	**132** (1)	62 (13, 6)	109 (13)	—	69 (4, 1)	117 (14)

E.2 TERMINATION CONTROL: AN ANSWER BUDGET

Between 10 and 27 percent of greedy AIME traces at 4B, and 3 to 23 percent at 8B, never emit `</think>` and are cut at the token limit with no answer. To separate a termination effect from a reasoning effect we reran greedy and conflict steering with an *answer budget*: if the thinking block is still open at $32{,}768 - 2{,}048$ tokens, `</think>` is forced and 2,048 tokens are reserved for the answer (a form of budget forcing (Muennighoff et al., 2025)). Closed greedy answers have a median length near 1,000 tokens. With the budget, at 8B: greedy 66.67 (60/90, 0.0) and GPQA 58.08 (115/198); conflict steering 72.22 (65/90, 0.0) and 55.05 (109/198); flips 10↔5 on AIME. At 4B: greedy 78.89 (71/90) and 66.67 (132/198); conflict steering 78.89 (71/90) and 64.65 (128/198); flips 4↔4. Of the greedy traces that reached the budget, 3 of 16 (8B) and 4 of 19 (4B) became correct once forced to answer, against 0 of 14 and 1 of 16 of the steering traces. When a steered trace reaches the limit, forcing an answer rarely recovers it. A larger budget is the other natural repair: continuing each of the 19 open 4B greedy AIME traces from its stored 32,768 tokens with a fresh budget of the same size closes 18 of them after 9 to 13 thousand further tokens on average, but only 6 become correct, against 8 for the steered rerun of the same 19 problems. At 4B the open traces wander rather than loop, so more budget ends them without repairing them. This agrees with the natural-closure analysis in Table 4.

E.3 LOOP EXAMPLE AND PROBLEM DIFFICULTY

Problem difficulty. We group AIME problems by how many of the eight chain-of-thought samples solve them. At 4B, the net gain occurs on problems solved two to five times (6↔1). No answers

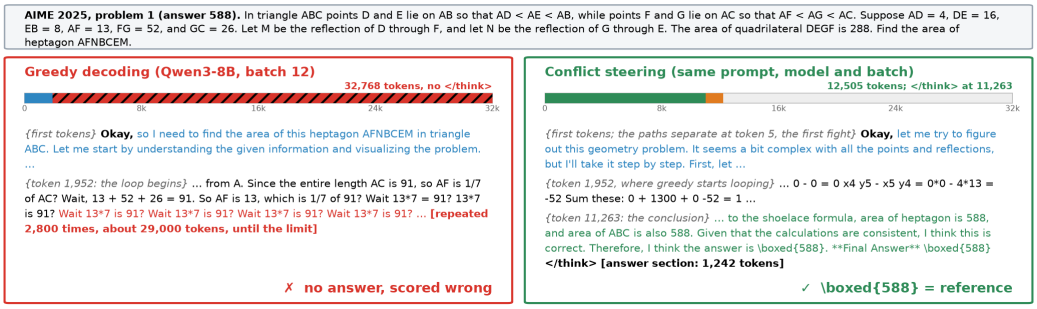

Figure 8: One paired decode of Qwen3-8B on AIME 2025 problem 1 (batch 12): greedy loops from about token 2,000, repeating one question 2,800 times until the limit. The steered decode leaves the greedy path at its fifth token, closes after 11,263 tokens and answers 588, the reference.

change on the 49 problems solved by all eight samples or the 12 solved at most once. At 8B, the net gain instead occurs on problems solved by all eight samples (4↔0), where greedy loses otherwise easy problems to loops. At 32B (AIME 2024 and 2025) the flips are spread over the bins (7↔5).

F Efficiency

Steering adds $M-1$ products with the unembedding matrix per decoding step while sharing one read of that matrix. Measured on the paired second-draw cells at the same batch size, our unoptimized research fork generates 225 against 452 tokens per second at 4B, 220 against 434 at 8B and 120 against 152 at 32B (steering against greedy), that is 0.50, 0.51 and 0.79 of greedy throughput.

Conflict steering has no stopping rule of its own. The thinking block ends when the model emits `</think>`, as in greedy decoding. Table 14 reports the mean number of generated tokens per problem for every method on AIME-3yr and the pooled code sets. Conflict steering is within ten percent of greedy at every scale, shorter at 4B (3.7 percent fewer tokens on AIME-3yr) and slightly longer at 8B (6.3 percent more), and on code it is 3 to 15 percent shorter. The Soft Thinking baselines are shorter everywhere, which is what the cold stop is designed to do, and they are also less accurate on AIME.

Table 14: Generation length: mean generated tokens (thinking and answer) per problem for the deterministic rows and per sample for the @8 rows, first paired draws. The steering fallback counts the greedy decode plus the steered rerun of the traces greedy leaves open. Dashes: not run. Shaded rows: ours.

	AIME 2024–26				Code (700)			
Method	4B	8B	14B	32B	4B	8B	14B	32B
Greedy	20,758	17,840	16,856	14,329	9,266	10,328	7,516	7,446
CLD hard (commit v_I)	19,916	18,344	—	15,132	8,955	10,487	—	—
Conflict steering	19,991	17,072	16,331	14,552	8,924	9,475	7,281	7,464
Steering fallback	26,985	22,230	18,679	16,081	10,696	14,073	8,484	8,000
CoT @8	20,281	16,040	14,958	14,154	8,972	7,917	—	—
Soft Thinking @8	19,355	15,047	—	13,509	—	—	—	—
ST+Gumbel @8	19,354	16,208	—	14,010	—	—	—	—
Conflict steering @8	20,217	17,353	—	—	—	—	—	—
Conflict steering, Em @8	20,378	17,001	—	14,310	8,844	8,533	—	—

G Task-dependent structure of late-layer conflict

Fight rate is a task signature, not a success predictor. Over the first 2,000 generated tokens, median fight rates are 0.031 and 0.050 on AIME at 4B and 8B, and 0.045 and 0.085 on GPQA (Figure 9). The per-problem distributions of AIME and GPQA barely overlap at 8B (AUROC 0.97) and overlap more at 4B, and the Otsu threshold of the pooled histogram (Otsu, 1979) (0.070 at

8B, 0.040 at 4B) separates them with no labels. Inside GPQA the same statistic is at chance for predicting whether steering helps or hurts a problem. The five contest sets, where steering does not help, have AIME-like rates (0.028 to 0.031 at 4B and 0.041 to 0.049 at 8B, with 40 to 42 and 29 to 31 percent of fights below τ, against 0.030 and 0.044 with 41 and 30 percent on AIME), and code has 0.037 to 0.072. Fight frequency alone therefore does not explain whether steering helps.

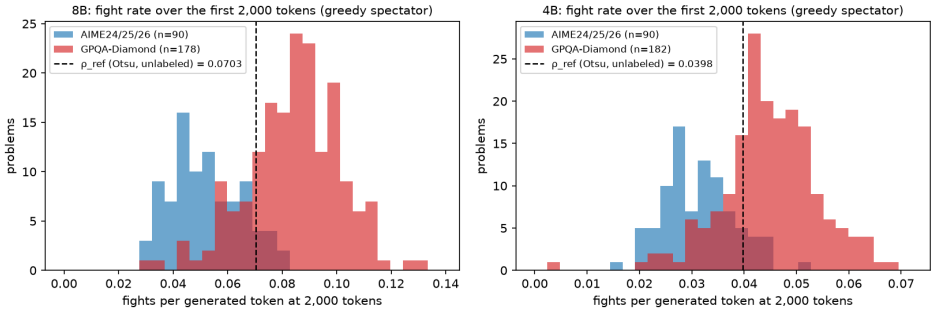

Figure 9: Per-problem fight rate over the first 2,000 thinking tokens, greedy spectator decode, AIME against GPQA-Diamond, with the unlabeled Otsu threshold. Contest math produces fewer fights than graduate science questions at both scales.

The content of the conflict differs. Figure 10 classes the intermediate proposals at near fights. On 8B GPQA the trough proposes *perhaps* 42 times per thousand near fights against 19 on AIME, and *Wait* 48 against 29. On AIME it more often proposes step-closing structure: paragraph breaks 43 times against 12, and *Therefore* 21 against 5. A fight-rate router and two token-based gates did not improve results, because neither feature predicts whether following the intermediate layer helps within a task, so neither is used as a routing signal. Late-layer conflict thus reflects task-dependent reasoning dynamics: simple statistics reveal the task difference but do not decide when an individual conflict should be followed, which remains open.

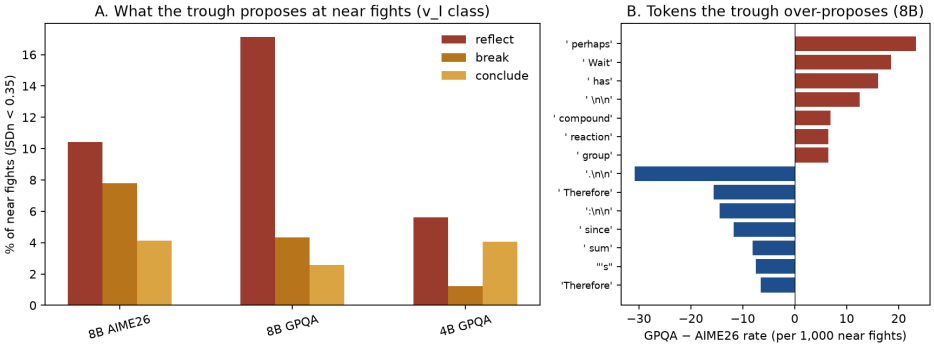

Figure 10: Left: share of near fights per token class of the intermediate proposal, per spectator set. Right: the tokens the intermediate layer over-proposes on GPQA relative to AIME 2026 (positive) and the reverse (negative), per thousand near fights, 8B.

H Reproducibility Notes

Method settings. Table 15 gives the configuration behind every conflict-steering cell in the paper: one value per scale for the model and its band, one shared value for everything else. Layer indices are 0-based and F is the final block. The band is the set of layers decoded with the final norm and unembedding (Section 3.1), and a fight is defined in Section 3.2. The sampled rows apply to every @8 cell of Table 12, ours and the baselines alike. The presence-penalty arms use $\beta \in \{0.5, 1.0\}$ as described in Appendix E.1.

Table 15: Conflict-steering configuration. Top: values that depend on the scale. Bottom: values shared by all scales. Layer indices are 0-based. The sampled rows apply to every @8 cell.

Setting	4B	8B	14B	32B
Model	Qwen3-4B-Thinking-2507	Qwen3-8B	Qwen3-14B	Qwen3-32B
Transformer blocks L	36	36	40	64
Final block F	35	35	39	63
Band width M	5	4	4	3
Band layers	31–35	32–35	36–39	61–63
Top-k, sampled cells	20	20	20	30
Entropy-matching support (Eq. 4)	20	20	20	30
Threshold τ	0.35			
Steepness k	14			
λ_{near}	1.0			
λ_{far}	0.0			
Temperature, greedy cells	0			
Temperature, sampled cells	0.6			
Top-p, sampled cells	0.95			
Min-p, sampled cells	0			
Samples per problem, sampled cells	8			
Maximum new tokens	32,768			
Precision	bf16			
Random seed	0			

Serving configuration. Greedy decoding is deterministic only within one serving configuration (Section 5.1), so Table 16 records the batch size, the maximum number of concurrently running requests, of every cell family. Each comparison in the paper pairs its two arms at one of these values. The KV-cache memory fraction is 0.90 at 4B and 0.92 at 8B and 32B.

Table 16: Serving configuration per cell family: batch size, the maximum number of concurrently running requests. Both arms of every comparison run at the same value.

Cells	4B	8B	14B	32B
AIME 2024–2026	20	12	30	16
GPQA-Diamond	24	28	48	15
AMC23	40	24	—	32
MATH-500	30	32	—	28
GSM8K	48	48	—	40
Five contest sets	20	12	—	16
HumanEval, MBPP, LiveCodeBench	16	16	32, 32, 24	16
Second paired draw	16	16	—	12

Hardware and software. All runs use NVIDIA A100-SXM4-80GB GPUs: one GPU at 4B and 8B, and two GPUs with tensor parallelism 2 at 32B. The serving stack is our fork of the Soft Thinking implementation (Zhang et al., 2025b), built on SGLang 0.4.6.post1 (Zheng et al., 2024), PyTorch 2.6.0 and CUDA 12.4. Steering runs inside the model forward pass and emits ordinary tokens. The presence penalty uses SGLang's stock implementation (Appendix E.1). Each experiment result records the host, GPU, driver, tensor-parallel size, batch size, memory fraction and code commit, together with fight counts, interpolation fractions, the d histogram and fight entropies.

Scoring. Mathematics and GPQA-Diamond are scored by a rule-based match of the final answer extracted from the completion against the reference (numeric equivalence for the mathematics sets, the selected option for GPQA-Diamond). Code is scored by executing the fenced solution block against the benchmark's test suite (Appendix B.3). Traces that never emit `</think>` within the token limit are scored on whatever final answer can be extracted from them and are reported separately as the no-`</think>` percentage.