TOKEN CAPS, NOT PROBES: MANAGING NON-TERMINATION IN DISTILLED REASONING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how thinking-token budgets relate to accuracy in DeepSeek-R1-Distill-Qwen-7B on GSM8K, MATH-500, and AIME. On GSM8K, accuracy peaks at a very small forced thinking budget, while on MATH-500 short budgets remain well below uncapped accuracy. On AIME, generations divide by whether the model ends its reasoning within a 10,000-token ceiling. Generations that terminate are usually correct; nearly half never terminate and consume most of the inference compute. Whether a problem terminates changes for about a quarter of problems under temperature sampling or across hardware. Forcing termination at a fixed cap requires no predictor and traces a compute–accuracy frontier: at 4,000 tokens it saves about 39% of inference compute with accuracy no lower than generating to the 10,000-token ceiling. Linear probes on early hidden states do not predict non-termination at the primary evaluation point; one position in a nine-position sweep (token 125) is at the threshold for significance after correction for multiple testing; replication is needed. We also show that the $\max(\mathrm{AUC}, 1-\mathrm{AUC})$ correction, valid for unsupervised probes, inflates the performance of supervised ones and produced a spurious positive result in our initial analysis. These results suggest that token-count caps, rather than activation probes, are the practical lever for reducing inference cost in this setting.

1 INTRODUCTION

The introduction of chain-of-thought prompting (Wei et al., 2022) and extended-thinking models such as OpenAI o1 (OpenAI, 2024) and DeepSeek-R1 (DeepSeek-AI, 2025) has restructured the computational cost of language model inference. These systems invest the majority of their token budget in long internal monologues before emitting a final answer, motivated by mounting evidence that test-time compute scaling improves reasoning quality (Snell et al., 2024; Wu et al., 2024; Brown et al., 2024).

The assumption that more thinking uniformly helps has received limited scrutiny at the level of individual generations. Aggregate benchmarks confirm improvements on average, but the per-problem distribution of token utility is poorly understood. A generation that converges cleanly at 4,000 tokens and one that loops for 10,000 tokens without terminating consume comparable compute yet exhibit radically different accuracy profiles. If these two regimes are separable—either by problem identity or by early generation signals—they define distinct optimization targets.

We investigate this question through a systematic study of DeepSeek-R1-Distill-Qwen-7B (and its 1.5B variant) across three benchmarks of increasing difficulty: GSM8K (Cobbe et al., 2021), MATH-500 (Hendrycks et al., 2021), and AIME 1983–2024 (Neubig, 2024). Our experimental program addresses four questions: (1) What is the minimum thinking budget at which accuracy stabilizes? (2) Does non-termination track difficulty? (3) Is non-termination stable under decoding? (4) Can internal representations predict convergence failure?

Findings. For (1), GSM8K budget-forced accuracy peaks at 77.8% (B=64), exceeding uncapped accuracy (67.0%); MATH-500 reaches 51.4% at B=512 vs. 61.4% uncapped; AIME has no finite B^* within 10,000 tokens. For (2), non-converged accuracy is 0.0% by evaluation rule; convergence

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.

rate decreases monotonically with era (68.1%, 50.0%, 40.9%). For (3), temperature sampling raises the aggregate convergence rate from 60.5% (greedy) to 72.4% (mean per-seed), and 23.5% of individual problems change convergence status between decoding modes. For (4), after correcting a $\max(AUC, 1 - AUC)$ bias, probe AUC at layer 20, token 150 is indistinguishable from chance (0.488, $p = 0.576$); a nine-position sweep finds token 125 at the threshold for significance after Bonferroni correction (AUC $= 0.646$, p=0.0049, threshold p<0.0056) and token 175 nominally significant (p=0.014) but not surviving correction.

Contributions.

- Sub-512 budget forcing on GSM8K and MATH-500: GSM8K accuracy peaks at 77.8% (B=64), exceeding uncapped (67.0%); MATH-500 reaches 51.4% at B=512 vs. 61.4% uncapped, indicating a higher saturation point.
- Termination on AIME is sensitive to decoding and hardware: cross-environment flip rate 24.0% (κ=0.504); greedy-vs-temperature item-level flip rate 23.5% (κ=0.465, 58.0% stably convergent), attributing variation to knife-edge fp16 sensitivity.
- Characterizing the early-exit token economy: oracle savings 65.0%; budget-forced cap at C=4,000 saves 39.0% of compute with accuracy 2.0 pp above uncapped (46.5% vs. 44.5%).
- Identifying and correcting a $\max(AUC, 1 - AUC)$ artifact in supervised probe evaluation; token 125 is at the threshold for significance after Bonferroni correction (AUC $= 0.646$, p=0.0049, threshold p<0.0056); token 175 is nominally significant (p=0.014) but does not survive; the design has 80% power to detect AUC ≥ 0.65 (permutation null simulation, null SD $= 0.060$); the primary evaluation (layer 20, token 150) shows AUC $= 0.488$.
- Documenting that post-cutoff AIME 2025 results (11/11 converged correct, 0/19 non-converged by evaluation rule) are consistent with the historical convergence–correctness pattern.

2 Related Work

Test-time compute scaling and chain-of-thought. Chain-of-thought prompting (Wei et al., 2022) showed that intermediate reasoning steps improve multi-step problem solving; Kojima et al. (2022) extended this to zero-shot settings. Self-consistency (Wang et al., 2023) and tree-of-thoughts (Yao et al., 2024) aggregate over multiple reasoning paths to improve reliability. Snell et al. (2024) showed that test-time compute exhibits diminishing returns at the aggregate level; our work extends this by showing that on hard benchmarks, many generations fail to terminate and dominate compute. Wu et al. (2024) and Brown et al. (2024) characterize inference compute scaling laws; Muennighoff et al. (2025) introduced budget forcing in the s1 framework, which we adopt and extend. Nikankin et al. (2024) find models solve arithmetic through heuristic shortcuts, which may partially explain knife-edge convergence.

Inference efficiency and early exit. Elbayad et al. (2020) and Schuster et al. (2022) apply early-exit ideas to transformer depth; Leviathan et al. (2023) reduce latency via speculative decoding; Schwartz et al. (2020) match model size to instance complexity. Our fixed-cap policy is the simplest instance of this class applied to token budget allocation in chain-of-thought.

Reasoning failure modes. Pipis et al. (2025) document a looping failure mode specific to R1-distilled models: students loop substantially more than their teacher. Chen et al. (2024) study unnecessary token production in eventually-correct generations. Bansal et al. (2024) show that compute-optimal sampling from weaker models can outperform larger models; our 1.5B results are consistent with their framework.

Probing internal representations. Linear probing is a standard tool for auditing neural network representations (Belinkov, 2022; Hewitt & Manning, 2019; Tenney et al., 2019). Gurnee et al. (2023) locate task-specific neurons with sparse probing; Burns et al. (2023) and Li et al. (2024) probe for truthfulness and apply interventions. Our probe application—predicting eventual generation outcomes from early hidden states—differs from prior work in that the label is a property of the full

Table 1: Budget-forced accuracy at sub-512 budgets and uncapped accuracy (7B, greedy). GSM8K peaks at B=64 (77.8%, exceeding uncapped 67.0%); MATH-500 reaches 83% of uncapped at B=256, indicating a higher saturation point. AIME yields no finite B^* within 10,000 tokens.

Benchmark	Acc@32	Acc@64	Acc@128	Acc@256	Acc@512	Uncapped	% Uncapped
GSM8K	77.4%	77.8%	74.8%	65.6%	67.2%	67.0%	98%
MATH-500	40.0%	45.4%	48.0%	51.2%	51.4%	61.4%	83%
AIME	4.0%	5.5%	6.5%	—	—	44.5%	—

trajectory. The $\max(AUC, 1 - AUC)$ correction appropriate for Burns et al. (2023)'s unsupervised probes is not appropriate here; we document the distinction explicitly.

Knowledge distillation and scaling. DeepSeek-R1's distillation (Hinton et al., 2015) trains on reasoning traces from a stronger teacher; high non-termination rates may reflect the student imitating successful traces without generalizing the convergence mechanism. Kaplan et al. (2020) and Hoffmann et al. (2022) established pretraining scaling laws; the 7B-to-1.5B convergence gap (53.5% vs. 31%) is consistent with a threshold effect in self-termination capability. The DeepSeek-R1 family (DeepSeek-AI, 2025) builds on Qwen2 (Yang et al., 2024), fine-tuned on R1 reasoning traces (Ouyang et al., 2022).

3 Experimental Setup

Models. DeepSeek-R1-Distill-Qwen-7B (28 layers, hidden size 3,584) and DeepSeek-R1-Distill-Qwen-1.5B (28 layers, hidden size 1,536), both loaded in float16 on an NVIDIA RTX A5000 (24 GB VRAM).

Benchmarks. **GSM8K** (Cobbe et al., 2021): 500 problems from the test split (seed 42). Grade-school arithmetic. **MATH-500** (Hendrycks et al., 2021): 500 problems (full curated test set). Competition-level mathematics. **AIME 1983–2024** (Neubig, 2024): 200 problems from the 933-problem corpus (seed 0). Olympiad-level. Nondeterminism experiments use seed 42 (noted where applicable).

Convergence definition. A generation is *converged* if it emits `</think>` before the 10,000-token ceiling. The position at which `</think>` appears is t_{end}.

Budget-forcing mechanism. A custom `LogitsProcessor` monitors thinking-phase token count. When the counter reaches budget B, the processor forces `</think>` by setting all other logits to $-\infty$. The processor initializes in thinking mode because the chat template opens the thinking phase prior to the first generated token.

Minimum sufficient budget. B^* is the smallest B at which accuracy achieves $\geq$95% of uncapped accuracy, sustained across all larger budgets. Confidence intervals use bootstrap resampling (1,000 iterations); proportions use Wilson score intervals.

4 Budget Saturation

4.1 Early saturation on GSM8K and MATH-500

Table 1 presents budget-forcing results at 32, 64, 128, 256, and 512 tokens alongside uncapped accuracy.

GSM8K budget-forced accuracy peaks at B=64 (77.8%), exceeding uncapped accuracy (67.0%), then declines — a non-monotone pattern consistent with an overthinking effect where forced short thinking elicits direct answers, while uncapped thinking leads to answer drift (Chen et al., 2024). MATH-500 shows a qualitatively different profile: accuracy rises monotonically from 40.0% at

Table 2: AIME convergence split (7B, greedy, uncapped, n=200). [a]By evaluation rule: no \boxed{} emitted within ceiling. [b]10,000 thinking tokens; answer not generated in uncapped run.

Group	n	Share	Accuracy	Avg. Tokens
Converged	107	53.5%	83.2%	≈4,675
Non-converged	93	46.5%	0.0%[a]	10,000[b]
Overall	200	—	44.5%	≈7,150

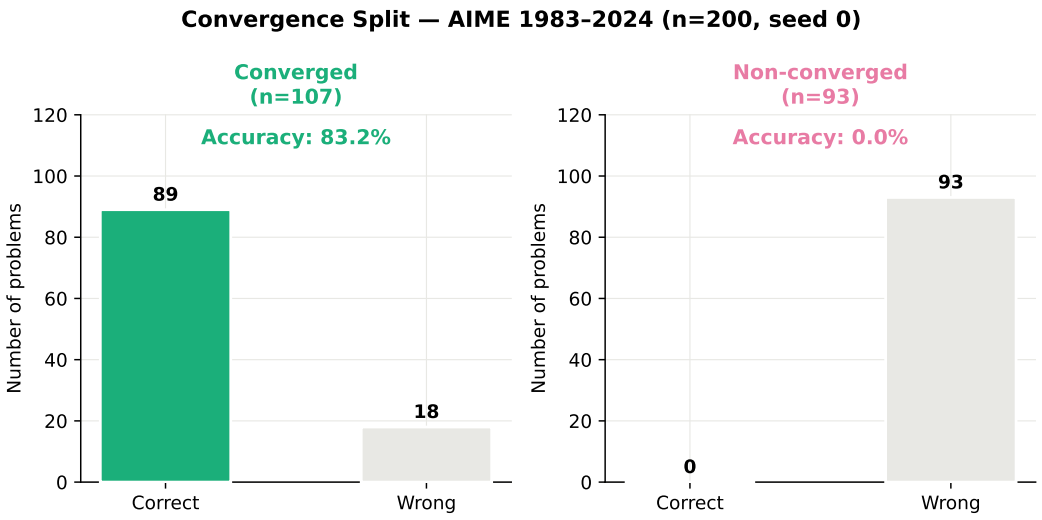


Figure 1: AIME convergence split (7B, greedy, seed 0, n=200). Converged generations (n=107) achieve 83.2% accuracy; non-converged generations (n=93) exhaust the 10,000-token ceiling with 0.0% accuracy by evaluation rule (no \boxed{} emitted).

B=32 to 51.4% at B=512 (83% of uncapped at B=256), indicating competition-level mathematics requires substantially longer reasoning chains.

4.2 Non-termination on AIME

AIME reveals a qualitatively different structure. Generations split by whether the model emits </think> before the token ceiling; converged generations achieve 83.2% accuracy, non-converged are scored 0.0% by the evaluation rule (no boxed answer emitted). Table 2 and Figure 1 report the primary analysis (convergence rate 53.5%, n=200, seed 0). The 0.0% is a scoring artifact: the evaluation rule never generates an answer for non-converged problems, so any correct reasoning in the thinking phase goes unscored.

Natural convergence budget curve. Figure 2 shows accuracy and convergence rate as a function of natural t_{end} threshold. Essentially no problem converges naturally below 1,024 tokens (<0.5% convergence rate). Convergence accumulates slowly: 24.5% at B=4000, 42.5% at B=6000, 53.5% at the ceiling. No saturation point exists within 10,000 tokens.

4.3 Non-termination is not explained by difficulty alone

Table 3 breaks down results by exam era, using publication year as an externally motivated difficulty proxy: the AIME has documented increasing difficulty over its four decades (Neubig, 2024), and year is independent of model outputs. Problems are partitioned by publication year into three groups: Early (≤2003, n=72), Mid (2004–2014, n=62), and Recent (≥2015, n=66).

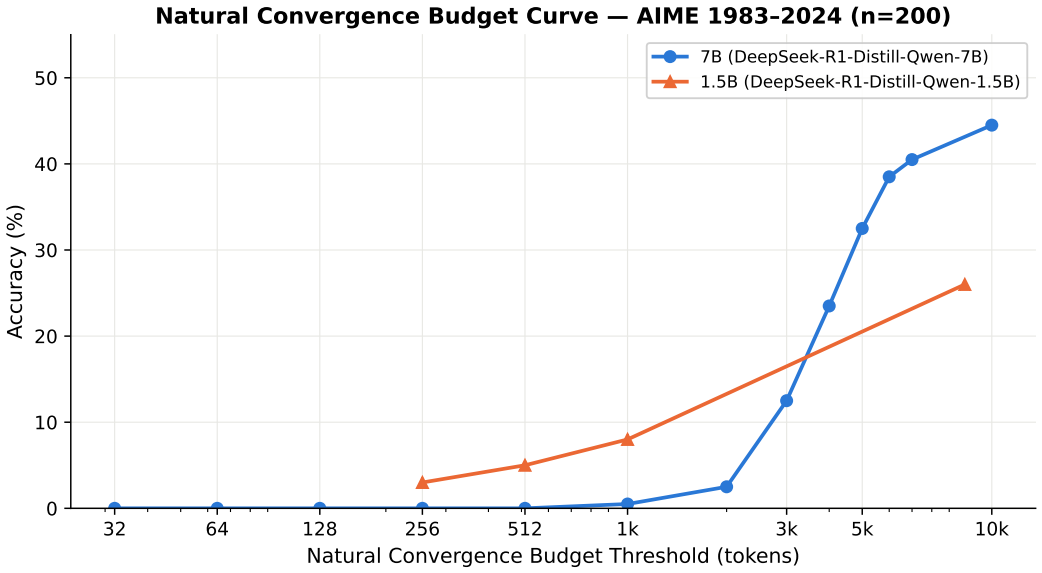


Figure 2: Accuracy as a function of natural convergence budget threshold on AIME (7B and 1.5B, greedy). The 7B curve ends at 44.5% (uncapped); the 1.5B curve ends at 26.0% (the 1.5B points below 10,000 tokens are budget-forced). The absence of a saturation point distinguishes AIME from GSM8K and MATH-500.

Table 3: AIME accuracy by convergence status and exam era (n=200, greedy, year-based groups). Non-converged accuracy is 0.0% in every era by evaluation rule (no `\boxed{}` emitted within the 10,000-token ceiling); convergence rate decreases monotonically with era.

Era	n	**Conv. Rate**	**Conv. Acc.**	**Non-conv. Acc.**
Early ($\leq$2003)	72	68.1%	87.8%	0.0%
Mid (2004–2014)	62	50.0%	87.1%	0.0%
Recent ($\geq$2015)	66	40.9%	70.4%	0.0%
All	200	53.5%	83.2%	0.0%

Convergence rate decreases monotonically with era (68.1%, 50.0%, 40.9%), but non-converged accuracy is 0.0% in every era by evaluation rule. Converged accuracy is 87.8%, 87.1%, and 70.4% across eras. Even in the early era, 31.9% of generations never converge, consistent with non-convergence not being reducible to problem difficulty alone.

4.4 Scaling: 1.5B vs. 7B

Table 4 compares natural-convergence results for the 1.5B model on AIME.

The 1.5B model converges naturally in 31% of AIME instances; forced convergence at small budgets achieves 3.0%, 5.0%, 8.0% accuracy at $B \in \{256, 512, 1024\}$. Among the 31% natural-convergence instances, per-problem accuracy is 83.9%, consistent with the 7B pattern.

5 Convergence Is Sensitive to Decoding

5.1 Run-to-run numerical nondeterminism

Across three independent greedy-decoding runs on the same 200 AIME problems (seed 42, distinct from the seed-0 primary sample)—an initial run and two new independent runs—convergence rates are 57.5%, 60.5%, and 60.5% respectively. Greedy decoding is deterministic given identical

Table 4: Scaling comparison on AIME (natural convergence, 10,000-token ceiling, ANSWER_CAP=1,000). The 1.5B-to-7B gap is primarily a convergence-rate loss; per-problem accuracy among converged instances is preserved.

Model	Conv. Rate	Overall Acc.	Conv. Acc.
7B	53.5%	44.5%	83.2%
1.5B	31.0%	26.0%	83.9%

Table 5: Item-level convergence transitions between greedy decoding and temperature majority (T=0.6, k=4 seeds, n=200). Cohen's κ=0.465 (greedy vs. temperature majority).

Convergence Status	Problems	Share
Stable converged	116	58.0%
Stable non-converged	37	18.5%
Non-conv. $\rightarrow$ conv. (flip)	42	21.0%
Conv. $\rightarrow$ non-conv. (flip)	5	2.5%
Total flips	47	23.5%

inputs; inter-run variation arises from fp16 accumulation-order nondeterminism, differences in attention backend (SDPA vs. eager), and presence or absence of a LogitsProcessor modifying the logit computation path. We present this variation as a substantive finding: sub-percent floating-point differences are sufficient to flip individual problems between convergence and non-convergence, consistent with the knife-edge dynamical characterization developed below. The convergence-rate structure is preserved across all three runs.

To quantify cross-environment instability, we ran two additional greedy passes (Runs 2 and 3) on a separate compute node; both show flip rate= 0.240 (48/200) against Run 1 (Cohen's κ=0.504). Runs 2 and 3 on the same hardware are fully deterministic (flip rate= 0.000). The 24.0% cross-environment greedy flip rate matches the greedy-to-temperature flip rate (23.5%), indicating fp16 nondeterminism alone destabilizes convergence at a similar rate to temperature sampling.

5.2 Temperature validation

DeepSeek's documentation advises against greedy decoding for R1-distilled models, citing risk of repetitive generation (DeepSeek-AI, 2025). We test whether the termination split is a greedy (argmax) artifact by running the same 200 AIME problems at temperature T=0.6 across four seeds (0, 1, 2, 3) and applying majority vote per problem.

Per-seed convergence rates are 73.0%, 72.5%, 70.5%, 73.5% (mean 72.4%). The majority-vote convergence rate is 79.0% (158/200). Temperature sampling raises the aggregate convergence rate; item-level convergence transitions are reported in Section 5.3.

5.3 Item-level convergence instability

Figure 3 and Table 5 report convergence stability at the problem level.

23.5% of problems (47/200) change their convergence label between decoding modes: 42 non-convergent $\rightarrow$ convergent, 5 in the reverse direction; 58.0% are stably convergent and 18.5% stably non-convergent (Cohen's κ=0.465). The asymmetric pattern is consistent with greedy decoding occasionally locking into looping attractors that stochastic sampling escapes.

6 Early-Exit Token Economy

Oracle savings. Non-converged generations consume 930,000 of 1,430,223 total tokens (93 runs $\times$ 10,000-token ceiling; the uncapped run never generates answer tokens for non-converged problems,

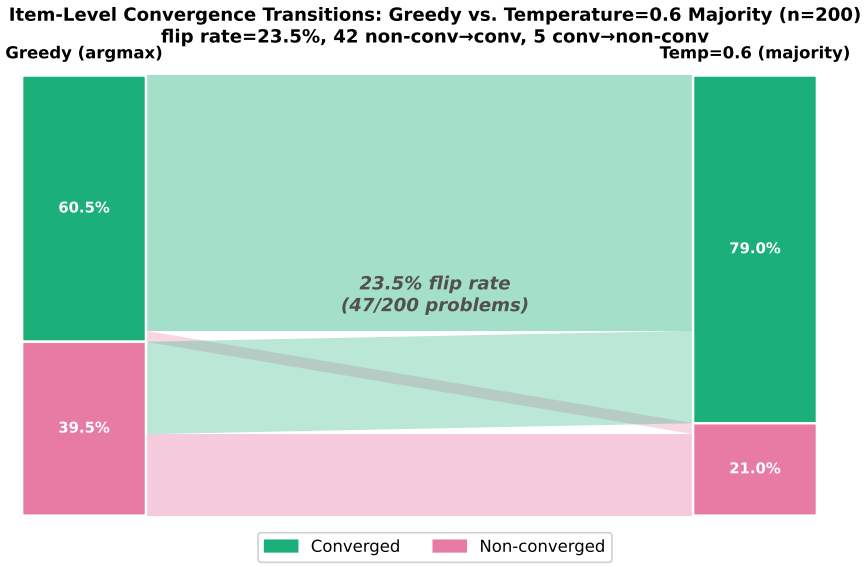


Figure 3: Item-level convergence transitions between greedy (argmax) and temperature majority decoding (T=0.6, k=4). Each bar segment represents a convergence-transition category. 23.5% of problems change status, predominantly from non-convergent to convergent under temperature sampling.

Table 6: Budget-forced fixed-cap policy (AIME, n=200, 7B). Compute saved is relative to uncapped total (1,430,223 tokens). Uncapped accuracy: 44.5%. Oracle ceiling: 65.0%.

Cap C	Compute Saved	Accuracy	Acc. Change	Nat. Conv.
2,000	59.7%	31.0%	−13.5 pp	6/200
3,000	48.6%	40.5%	−4.0 pp	26/200
4,000	39.0%	46.5%	+2.0 pp	49/200
5,000	30.5%	48.5%	+4.0 pp	71/200
6,000	22.9%	52.0%	+7.5 pp	85/200
7,000	15.4%	51.0%	+6.5 pp	90/200

so these are excluded from the token count). An oracle terminating non-converging generations at start saves 65.0% of total inference tokens — the theoretical ceiling for any early-termination policy.

Budget-forced fixed-cap policy. A practical policy requiring no prediction truncates all generations at cap C, forcing `</think>` at that point and collecting the best-available answer. Table 6 and Figure 4 report compute saved and accuracy retained across the Pareto frontier. C was selected on the evaluation set; held-out validation is deferred to future work.

At C=4,000, 39.0% of compute is saved with accuracy 2.0 pp *above* uncapped (46.5% vs. 44.5%); at C=6,000, accuracy peaks at 52.0% (7.5 pp above uncapped, 22.9% compute saved).

Answer-cap saturation caveat. Of the 151 forced generations at C=4,000, 98 (64.9%) exhaust the 1,000-token answer window without emitting `\boxed{}` and are scored 0; accuracy gains above uncapped arise from the remaining subset that do produce a boxed answer. Among all forced generations, accuracy is 19–32% (e.g., 30.5% at C=4,000): forced termination recovers real answers from looping trajectories, confirming that uncapped 0.0% reflects the scoring rule, not an intrinsic ceiling. For reference, MATH-500 at B=32 has a 60.2% answer-cap saturation rate (301/500).

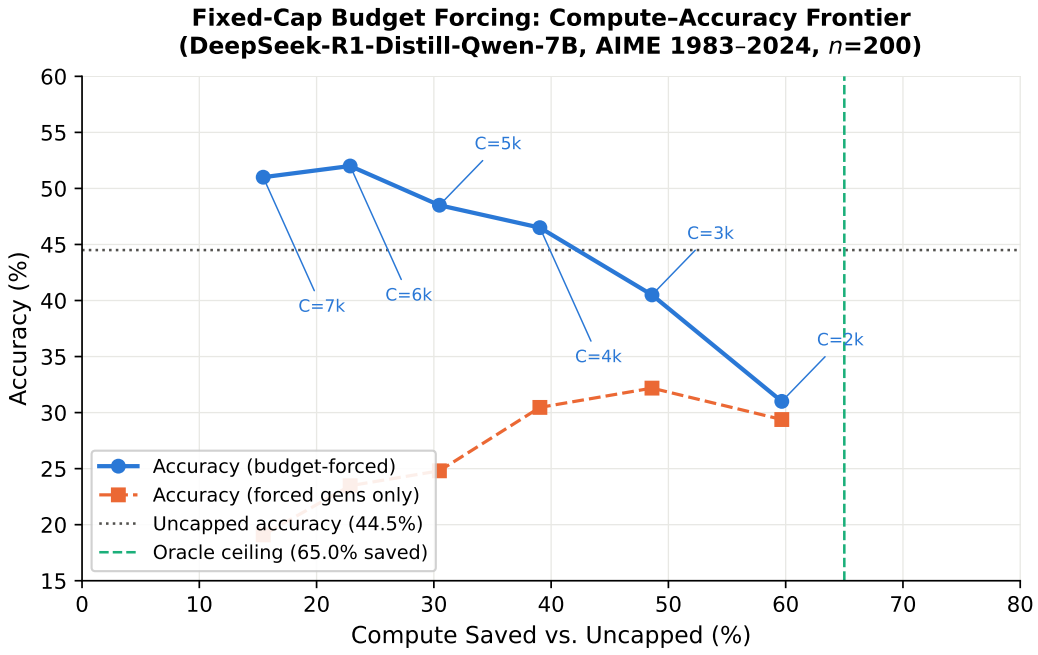


Figure 4: Compute–accuracy Pareto frontier for budget-forced fixed-cap inference on AIME (7B, greedy, n=200). Blue line: overall accuracy (all 200 problems). Orange dashed line: accuracy among forced generations only. Dotted horizontal line: uncapped accuracy (44.5%); dashed vertical line: oracle compute ceiling (65.0%). Caps above C=4,000 exceed uncapped accuracy owing to the 1,000-token answer window rescuing looping generations.

7 Activation Probes Show No Reliable Predictive Signal

7.1 Setup

We regenerate all 200 AIME problems (seed 42, Run A; convergence rate 57.5%, n_1=115 converged, n_2=85 non-converged) with forward hooks on all 28 layers at token position 150. A logistic regression probe (C=1.0, StandardScaler, liblinear) is evaluated via 5-fold stratified cross-validation with AUC-ROC; StandardScaler is fit on training folds only. Significance uses a generation-level permutation test (n=1,000) reassigning convergence labels across all 200 generations simultaneously. We additionally sweep nine token positions $\{50, 75, 100, 125, 150, 175, 200, 250, 300\}$ using layer-20 activations, applying Bonferroni correction for the nine simultaneous tests.

7.2 Methodological correction

Our initial analysis applied $\hat{A} = \max(AUC, 1 - AUC)$. This correction is appropriate for *unsupervised* probes under label-sign ambiguity (Burns et al., 2023), but invalid for supervised logistic regression: training labels fix the orientation, so no sign ambiguity exists. In high-dimensional settings (d=3,584, n=200), probes frequently achieve AUC < 0.5 by chance; the correction maps all such values above 0.5, inflating the signal. Our initial layer-20 AUC of 0.608 ± 0.080 used this invalid correction; we remove it below.

7.3 Corrected results

The generation-level permutation test at layer 20, token 150 yields observed AUC $= 0.488$ ($p = 0.576$; Figure 6 in Appendix F). The probe performs below the null mean. A sweep across selected layers at token 150 shows higher AUCs at earlier layers (layer 0: AUC$= 0.606$, p=0.027, nominally

significant but not corrected for multiple comparisons) with no layer surviving Bonferroni correction; see Table 11, Appendix D.

Token position sweep. Table 10 (Appendix D) reports the full nine-position sweep. Token 125 yields the highest observed AUC of 0.646 (p=0.0049, 10,000-permutation null), at the threshold for Bonferroni correction (threshold $p < 0.0056$ for nine tests). Token 175 is the second-highest at 0.626 (p=0.014) but does not survive correction. All remaining positions are consistent with chance ($p > 0.1$). The behavioral baseline shows no significant discrimination at any tested checkpoint (AUC range 0.48–0.55, all $p > 0.2$).

7.4 Bounded null result via power analysis

The permutation null has empirical SD $= 0.060$ (mean 0.501, n=1,000 permutations). Using a permutation-based power simulation with this null distribution, the minimum detectable AUC at 80% power is 0.65 (Figure 7, Appendix F). The observed interval is $0.488 \pm 1.96 \times 0.060 \approx [0.37, 0.61]$, consistent with no effect. Token 125 is at the threshold for Bonferroni correction (p=0.0049, threshold p<0.0056, 10,000-permutation null); token 175 does not survive (p=0.014). A C-sweep over $\{10^{-3}, \ldots, 10^{2}\}$ at token 150 is unchanged: best observed AUC 0.492 across regularization strengths, all $p > 0.5$.

8 Contamination Check on AIME 2025

To verify that results are not driven by training-set memorization, we run uncapped inference on all 30 post-cutoff AIME 2025 problems (released February 2025, strictly after the July 2024 training cutoff). The convergence–correctness pattern is consistent with the historical results: converged accuracy is 100% (11/11) and non-converged accuracy is 0% by evaluation rule (0/19); see Appendix G for full results.

9 Discussion and Conclusion

GSM8K accuracy peaks at B=64 (77.8%, exceeding uncapped 67.0%); MATH-500 has not saturated by 512 tokens; for AIME, a fixed cap at C=$4{,}000$ saves 39.0% of compute with accuracy 2 pp above uncapped (46.5% vs. 44.5%), requiring no predictor. Non-converged accuracy is 0.0% by evaluation rule (the pipeline never generates an answer for non-converged runs); the 23.5% flip rate and fp16 nondeterminism imply a knife-thin convergence boundary. The $\max(AUC, 1-AUC)$ correction inflates supervised probe performance (Burns et al., 2023).

Reproducibility Statement

Model weights are at `deepseek-ai/DeepSeek-R1-Distill-Qwen-7B` and `-1.5B` on HuggingFace; hyperparameters are in Section 3; code, scripts, and result files are in the anonymized repository. Seeds: AIME primary n=200 seed 0; nondeterminism/temperature n=200 seed 42 (seeds 0–3 for temperature); GSM8K 500 problems seed 42; MATH-500 complete test set.

AI Use Statement

We used AI writing tools for grammar and style assistance in preparing this manuscript. All scientific content, experimental design, data analysis, and conclusions are the authors' own work.

References

Hritik Bansal, Arian Hosseini, Rishabh Agarwal, Vinh Q Tran, and Mehran Kazemi. Smaller, weaker, yet better: Training LLM reasoners via compute-optimal sampling. In *Advances in Neural Information Processing Systems*, 2024.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT think that much for 2+3=? On the overthinking of o1-like LLMs. *arXiv preprint arXiv:2412.21187*, 2024.

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. In *Advances in Neural Information Processing Systems*, 2021.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. Depth-adaptive transformer. In *International Conference on Learning Representations*, 2020.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*, 2023.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems*, 2021.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4129–4138, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pp. 19274–19286, 2023.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 2024.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Graham Neubig. AIME 1983–2024 dataset. Technical report, Hugging Face, 2024. https://huggingface.co/datasets/gneubig/aime-1983-2024.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic without algorithms: Language models solve math with a bag of heuristics. *arXiv preprint arXiv:2410.21272*, 2024.

OpenAI. Learning to reason with LLMs. Technical report, OpenAI, 2024. URL https://openai.com/research/learning-to-reason-with-llms.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Charilaos Pipis, Shivam Garg, Vasilis Kontonis, Vaishnavi Shrivastava, Akshay Krishnamurthy, and Dimitris Papailiopoulos. Wait, wait, wait... why do reasoning models loop? *arXiv preprint arXiv:2512.12895*, 2025.

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*, 2022.

Roy Schwartz, Gabriel Stanovsky, Shmuel Swayamditta, Jesse Dodge, and Noah A Smith. The right tool for the job: Matching model and instance complexities. *arXiv preprint arXiv:2004.07453*, 2020.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

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

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

A Full Budget Sweep Results

Table 7: AIME natural convergence sweep (7B, greedy, n=200). Accuracy and convergence rate are defined by $t_{\text{end}} \leq B$.

Budget B	**Conv. Rate**	**Accuracy**
32	0.0%	0.0%
64	0.0%	0.0%
128	0.0%	0.0%
256	0.0%	0.0%
512	0.0%	0.0%
1,024	1.0%	0.5%
2,000	3.0%	2.5%
3,000	13.0%	12.5%
4,000	24.5%	23.5%
5,000	35.5%	32.5%
6,000	42.5%	38.5%
7,000	45.0%	40.5%
10,000	53.5%	44.5%

B Budget Forcing: Sub-Minimum Budgets

Table 8: Budget forcing at sub-minimum budgets (7B, AIME, n=200). All generations converge by construction.

Budget	**Conv. Rate**	**Accuracy**	**Mean Tokens**
32	100%	4.0%	232
64	100%	5.5%	264
128	100%	6.5%	328

C 1.5B Model Results

Table 9: 1.5B model budget-forced accuracy (AIME, n=200).

Budget	**Conv. Rate**	**Accuracy**	**Mean Tokens**
256	100%	3.0%	456
512	100%	5.0%	711
1,024	100%	8.0%	1,222
10,000 (natural)	31.0%	26.0%	8,285

D PROBE LAYER SWEEP AND TOKEN POSITION SWEEP: CORRECTED AUCS

Table 10: Token position sweep: layer-20 activation probe AUC across nine token positions (n=200, 5-fold CV observed AUC, 1,000-permutation full-sample null for all positions; token 125 and 175 rerun with 10,000 permutations, fixed C=1.0, liblinear solver, StandardScaler). Token 125 is at the Bonferroni threshold (p=0.0049, threshold $p<0.0056$, 10,000-permutation null; MC CI $[0.0035, 0.0063]$); token 175 is nominally significant (p=0.014) but does not survive.

Token	AUC	Permutation p	Note
50	0.539	0.261	
75	0.487	0.596	
100	0.520	0.358	
125	0.646	0.0049	at Bonferroni threshold[†]
150	0.487	0.576	primary evaluation point
175	0.626	0.014	nominally significant
200	0.533	0.305	
250	0.517	0.358	
300	0.563	0.129	

[†]p-value confirmed with 10,000-permutation null (MC error ± 0.0007, 95% CI $[0.0035, 0.0063]$); 1,000-permutation estimate was p=0.006. Token 125 sits at the Bonferroni threshold; replication is needed.

Table 11: Activation probe AUC across selected layers at token 150 (no $\max$ correction, n=200, 5-fold CV, fixed C=1.0, liblinear, 1,000-permutation full-sample null matching the primary evaluation). Layer 0 is nominally significant (p=0.027) but not corrected for multiple comparisons; no layer reaches Bonferroni threshold.

Layer	AUC	Permutation p
0	0.606	0.027
5	0.567	0.117
10	0.527	0.321
15	0.525	0.351
20	0.487	0.576
25	0.506	0.454
27	0.510	0.441

E SCALING TREND FIGURE

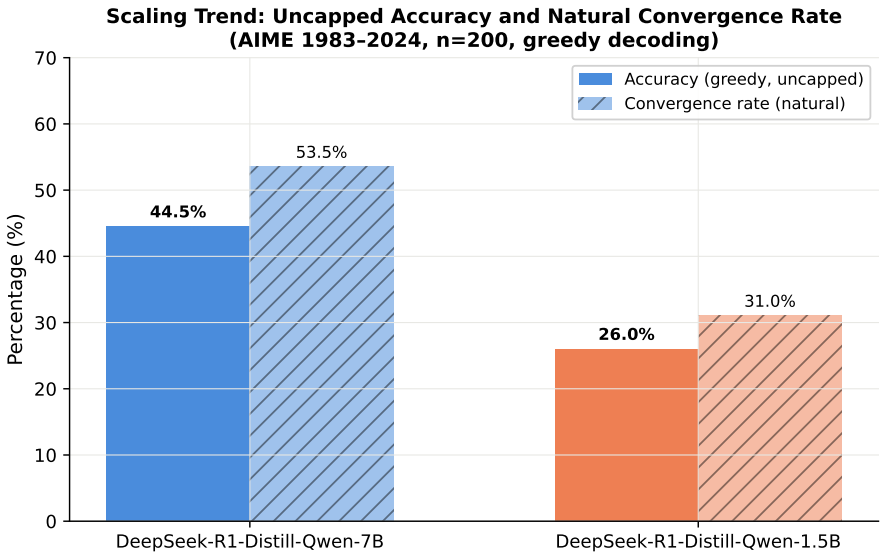


Figure 5: Uncapped accuracy and natural convergence rate for 7B and 1.5B models on AIME (greedy, n=200). The 7B–1.5B gap (53.5% vs. 31.0% convergence rate; 44.5% vs. 26.0% accuracy) is consistent with a threshold effect in self-termination capability.

F PROBE FIGURES: PERMUTATION TEST AND POWER ANALYSIS

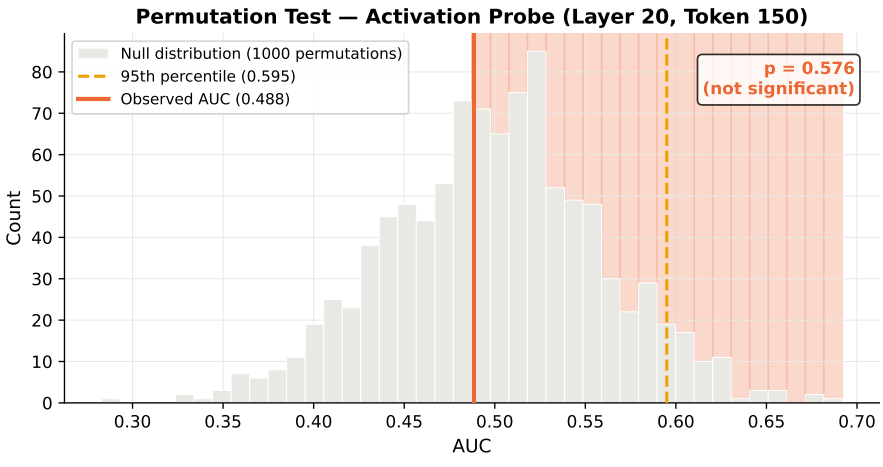


Figure 6: Null distribution of AUC under the generation-level permutation test (n=1,000, layer 20, token 150). Observed AUC $= 0.488$ (solid red line) falls within the null (mean 0.501, sd 0.060). $p = 0.576$.

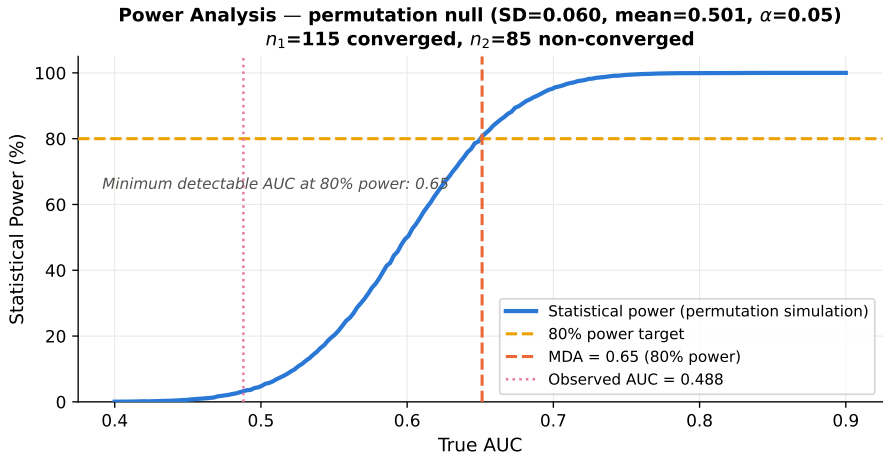


Figure 7: Statistical power as a function of true AUC (permutation-based simulation, null SD $= 0.060$, α=0.05). The minimum detectable AUC at 80% power is 0.65.

G Contamination Check on AIME 2025: Full Results

AIME 1983–2024 problems predate the training cutoff (July 2024); we test memorization by running uncapped inference on all 30 AIME 2025 problems (released February 2025, strictly post-cutoff).

Table 12: Contamination check: AIME 2025 (post-cutoff, n=30) vs. AIME 1983–2024 (n=200).

Metric	AIME 1983–2024	AIME 2025
Overall accuracy	44.5%	36.7%
Convergence rate	53.5%	36.7%
Converged accuracy	83.2%	100.0% (11/11)
Non-converged accuracy	0.0%	0.0% (0/19)

The overall accuracy drop (44.5% $\rightarrow$ 36.7%) is driven entirely by a lower convergence rate (53.5% $\rightarrow$ 36.7%), not by degraded reasoning quality among converged instances. Converged accuracy on 2025 problems is 100% (11/11)—higher than the 83.2% historical rate—and non-converged accuracy is 0% (0/19), consistent with the historical 0.0%. This is consistent with the convergence–correctness pattern holding on genuinely unseen data. Non-converged accuracy is 0.0% (0/19) by evaluation rule: the 10,000-token thinking ceiling is reached before `\boxed{}` is emitted, so no scored answer is generated — the same mechanism as in the primary seed-0 evaluation.