YOUR MODEL DIVERSITY, NOT METHOD, DETERMINES REASONING STRATEGY

Moulik Choraria * UIUC
Argyrios Gerogiannis UIUC
Anirban Das Capital One
Supriyo Chakraborty Capital One

Berkcan Kapusuzoglu, Chia-Hsuan Lee, Kartik Balasubramaniam
Capital One

Shi-Xiong Zhang, Sambit Sahu
Capital One

ABSTRACT

Compute scaling for LLM reasoning requires allocating budget between exploring solution approaches (*breadth*) and refining promising solutions (*depth*). Most methods implicitly trade off one for the other, yet why a given trade-off works remains unclear, and validation on a single model obscures the role of the model itself. We argue that **the optimal strategy depends on the model's *diversity profile*, the spread of probability mass across solution approaches, and that this must be characterized before any exploration strategy is adopted.** We formalize this through a theoretical framework decomposing reasoning uncertainty and derive conditions under which tree-style depth refinement outperforms parallel sampling. We validate it on Qwen-3 4B and Olmo-3 7B families, showing that lightweight signals suffice for depth-based refinement on low-diversity aligned models while yielding limited utility for high-diversity base models, which we hypothesize require stronger compensation for lower exploration coverage.

1 INTRODUCTION

Exploration via compute scaling has emerged as a powerful axis for improving LLM reasoning (Wei et al., 2023; Shao et al., 2024; Snell et al., 2024; Welleck et al., 2024). The idea is simple: sample N independent trajectories in parallel, then choose the best candidate (Cobbe et al., 2021; Wang et al., 2023); *tree search* methods take it further by iteratively expanding and refining the sampled trajectories via depth-based exploration (Zhao et al., 2023; Zhang et al., 2024). The landmark success of DeepSeek-R1 (Guo et al., 2025) in exploiting exploration during training, has spurred a wave of follow-ups on alternate strategies for training-time exploration, beyond basic parallel sampling (Xu et al., 2025; Zhuang et al., 2025; Yao et al., 2025). The template is standard: propose a strategy with some intuitive motivation, evaluate it on 1-2 chosen models, and compare against baselines which use parallel sampling such as GRPO/DAPO (Shao et al., 2024; Yu et al., 2025).

Despite the empirical gains, what is missing is an understanding of why a particular strategy works for a particular model. Current work rarely justifies choices like starting from instruct-tuned versus base checkpoints, or favoring parallel sampling over tree-based refinement. **We argue that these choices matter, and that any reasoning approach must first characterize its operational regime–its diversity profile, the spread of probability mass across distinct solution approaches–before adopting a strategy**. Without this, there is no principled basis for choosing: a strategy effective for one model class may fail on another because the underlying diversity regime differs.

To this end, we first develop a **theoretical framework** (§2) that decomposes reasoning uncertainty and derives conditions under which depth-based refinement outperforms parallel sampling as a function of the model's diversity profile. We then **empirically validate** (§3) these conditions across two

*Correspondence: moulikc2@illinois.edu; Part of the work was done during MC's internship at Capital One

model families, showing that the predicted regime boundaries hold and existing approaches are no longer as effective when applied outside their regime. While we do not claim that diversity profile is the sole factor–dataset choice, prompt design, and model scale can all confound the picture, as we discuss in (§4)–these results offer strong support for our position that the target model's diversity regime must be characterized before any exploration strategy is adopted.

2 FRAMEWORK

To study the role of model diversity, we first propose a framework to decompose reasoning uncertainty, in the same vein as Bakman et al. (2025). Our central question: for a fixed model/policy and compute budget, how should a strategy trade-off *breadth* versus *depth* exploration. At one extreme, standard parallel sampling represents one means of allocating all compute to breadth; on the other, depth-first refinement resembles tree search. We formalize this trade-off through an abstraction of Monte Carlo Tree Search (MCTS) (Świechowski et al., 2022).

Setup: Let q represent the problem of interest, and the model, parametrized via ϕ, generate reasoning trajectories as $\tau \sim p_\phi(\cdot)$. We define $\mathcal{T}$ as the set of all possible reasoning trajectories (for q), and assume that it admits a *partition* into m disjoint measurable sets $\mathcal{T} = \biguplus_{i=1}^{m} \mathcal{A}_i$. Here, each $\mathcal{A}_i$ intuitively represents a class of high-level solution approaches (e.g. analytical, induction-based etc.). To characterize the probability of finding a correct trajectory, we define:

$$\pi_i(\phi) = \mathbb{P}(\tau \in \mathcal{A}_i | \tau \sim p_\phi(\cdot)) \quad \text{(probability that sample belongs to an approach for model } \phi)$$

$$\theta_i(\phi, B) = \Pr(\text{correct} \mid \tau \in \mathcal{A}_i; \text{ budget } B). \quad \text{(model success rate under an approach)}$$

$$\bar{\theta}(\phi, B) = \textstyle\sum_i \pi_i \, \theta_i \quad \text{(marginal model success rate)}$$

Next, we define the the set of *viable* approaches $\mathcal{V} = \{i : \exists\, \tau \in \mathcal{A}_i, \tau \text{ is correct}\}$ i.e. approaches that contain at least one correct trajectory. We also define an ideal model ϕ^*, with the intuition being that it allocates masses to different approaches optimally via $\pi(\phi^*)$. We can now characterize the difficulty of the problem q by decomposing the reasoning uncertainty in into three components:
(i) *Aleatoric uncertainty:* $U_{\text{alea}}(q) = H[A \mid \phi^*]$, which represents the inherent difficulty of identifying the viable latent approaches $\mathcal{V}$, even for the ideal model. This stems from q and is irreducible.
(ii) *Epistemic breadth:* $U_{\text{epi}}^{\text{breadth}}(\phi) = D_{\text{KL}}(\pi(\phi^*) \,\|\, \pi(\phi))$, which measures the mismatch in approach selection for the given model against the ideal model. We hypothesize that reducing this term requires two stages: first, *diverse sampling* (breadth exploration) to increase coverage of approaches in the support of $\pi(\phi^*)$; then, *importance-weighted selection* among discovered approaches to correct the mismatch between $\pi(\phi)$ and $\pi(\phi^*)$.
(iii) *Epistemic depth:* $U_{\text{epi}}^{\text{depth}}(\phi, B) = \sum_{i \in \mathcal{V}} \pi_i(\phi^*) \, D_{\text{KL}}\big(\text{Bern}(\theta_i(\phi^*, B)) \,\|\, \text{Bern}(\theta_i(\phi, B))\big)$ i.e. the divergence in within-approach execution quality relative to the ideal model. Our hypothesis is that it complements breadth exploration, and can be reduced by *feedback and iterative refinement*. With this framework in place, we can now focus on designing an optimal exploration strategy.

An MCTS Allocation Strategy. Consider a budget of N trajectories. We split the budget into $N_e = \alpha N$ for *exploration* (discovering distinct approaches) and $N_f = (1-\alpha)N$ for *refinement* (improving within the most promising discovered approach i^*). This mirrors MCTS: N_e expands the tree across approaches while N_f concentrates on high-value branches. The question is: under what conditions does this outperform i.i.d. sampling?

Proposition 1. *Let $c_i := -\log(1-\theta_i)$ measure the per-sample refinement value of approach i, $c_{\text{rand}} := -\log(1-\bar{\theta})$ that of random sampling, and $R^{(t)} := c_{i^*}^{(t)} / c_{\text{rand}}$ the quality ratio of the best approach over the marginal. Then, assuming oracle identification the optimal approach i^*:*

(a) $\text{pass@}N_{\text{mcts}} \geq \text{pass@}N_{\text{rand}}$ *iff* $\alpha \leq 1 - 1/R^{(t)}$.

(b) Discovering i^* with probability $\geq 1-\eta$ under i.i.d. exploration requires $\alpha \geq \alpha_{\min}(\eta) := \frac{\log(1/\eta)}{N \log(1/(1-\pi_{i^*}))} \approx \frac{\log(1/\eta)}{N \pi_{i^*}}$.

Both conditions are jointly satisfiable iff $R^{(t)} \geq \frac{1}{1-\alpha_{\min}(\eta)}$.

The proposition reveals a tension between two competing constraints. Condition (a) imposes an upper bound on the exploration fraction: allocating too large an α diverts budget from refinement,

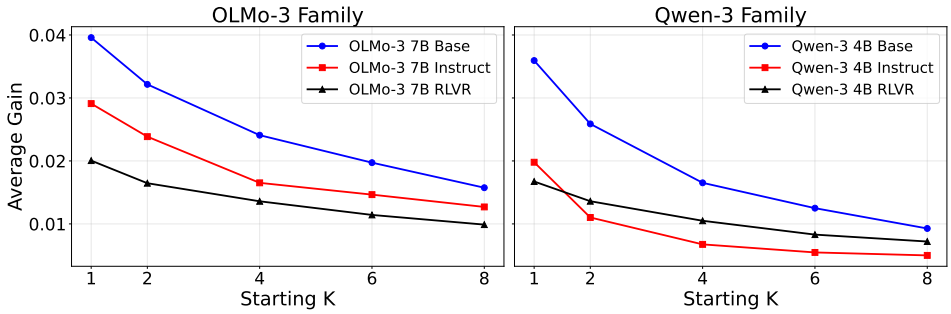

Figure 1: Average gain per additional rollout from budget K to 16. Base models sustain high gains at larger K, indicating unsaturated exploration; aligned variants saturate earlier, suggesting a more concentrated approach distribution.

where it would be more effective. Condition (b) imposes a lower bound: discovering rare approaches requires sufficient exploration. The two are jointly satisfiable only when the quality ratio $R^{(t)}$ is large enough to absorb the discovery cost. We note that this analysis assumes all refinement concentrates on a single best approach i^*, which represents the most favorable setting for depth-based strategies. Since MCTS cannot outperform random sampling under any weaker refinement allocation if it fails to do so here, the conditions in Proposition 1 are necessary for MCTS to be viable, and remain informative under this simplification. The proof is included in Appendix C.

3 Experiments

We evaluate across six models: OLMo-3 7B (Olmo et al., 2025) and Qwen-3 4B (Yang et al., 2025) on the base, instruct, and RLVR checkpoints. For each model and problem, we draw an i.i.d. pool of N=16 rollouts $\{\tau_1, \ldots, \tau_N\} \sim p_\phi(\cdot \mid q)$ on a 1024-problem subset of DeepMath (He et al., 2025) at hardest difficulty (level 9). The **Baseline** corresponds to standard i.i.d. sampling (e.g. GRPO) with a budget of N rollouts. We measure relative improvement via $\text{Lift}(N) = (\text{Pass@}N^{\text{method}} - \text{Pass@}N^{\text{i.i.d.}})/\text{Pass@}N^{\text{i.i.d.}}$, where negative lift indicates degradation.

The Model Gap. Before introducing our exploration strategies, we first characterize the diversity profiles of the six models under standard i.i.d. sampling. Figure 1 plots the average gain per additional rollout required to reach budget 16 from K, defined as $\text{Average Gain} = (\text{Pass@16} - \text{Pass@}K)/(16 - K)$ for $K \in \{1, 2, 4, 6, 8\}$. Base models sustain high average gains even at larger K, indicating that additional rollouts continue to uncover new viable approaches. Aligned variants, by contrast, saturate earlier: the marginal value of an extra rollout drops quickly, consistent with a more concentrated approach distribution. This gap motivates our central question: if base and aligned models occupy different diversity regimes, they should respond differently to breadth and depth exploration.

Stage 1: Breadth (Prefix Selection). We first generate N' short prefixes (< 256 tokens) and subselect N_e via a logprob-based diversity criterion (see Appendix B). This method accesses the full top-k logprob distribution and identifies promising prefixes, simulating an oracle that selects the N_e most distinct approaches, and requires no additional model calls beyond the N' generations. While there is no guarantee that our heuristics cluster prefixes into genuinely distinct approaches, one could bridge this gap with an external verifier for semantic clustering.

Stage 2: Depth (Refinement). Starting from the N_e selected prefixes, we allocate the remaining budget $N - N_e$ to within-branch refinement. We consider two variants:

(i) **ENT**: Following Hou et al. (2025), we identify high-entropy tokens along each trajectory and branch at these points, expanding the tree into alternative continuations.

(ii) **ENT+SR**: In addition to entropy-based branching, we incorporate iterative self-refinement (Madaan et al., 2023). When a rollout fails (determined by final-answer correctness), the same model generates a short critique (< 512 tokens) diagnosing where the reasoning went wrong. This feedback is then injected into subsequent expansions to discourage repeated failure patterns.

Table 1: Pass@16 for **Baseline**, **ENT** and **ENT+SR**. The Lift showcases the gain over the **Baseline**.

Model	Baseline	ENT	ENT+SR	ENT Lift	ENT+SR Lift
OLMo-3 7B base	0.8506	0.9072	0.8828	6.65%	3.79%
OLMo-3 7B instruct	0.6797	0.8154	0.8086	19.96%	18.96%
OLMo-3 7B RLVR	0.4385	0.6201	0.5967	41.41%	36.08%
Qwen-3 4B base	0.9004	0.6377	0.6406	-29.18%	-28.85%
Qwen-3 4B instruct	0.9121	0.9180	0.9170	0.65%	0.54%
Qwen-3 4B RLVR	0.3965	0.5059	0.6494	27.59%	63.78%

For ENT and ENT+SR, we fix the prefix-generation budget to $N' = 32$ and retain $N_e = 6$ prefixes for refinement, leaving the remaining rollout budget for depth exploration within the selected branches. We chose the prefix-selection rule using a validation sweep on Qwen-3 4B models over a disjoint held-out subset of DeepMath, and found that the logprob-based diversity criterions used in Appendix B gave the most reliable trade-off between broad initial coverage and downstream refinement quality. We then hold these hyperparameters fixed across all six models and for both methods.

Results. Table 1 confirms the prediction from the model gap analysis. Within each family, base models benefit least from our pipeline, while instruct and RLVR variants show progressively larger lifts. For OLMo, ENT lift increases from 6.65% (base) to 41.41% (RLVR), and ENT+SR follows the same trend. Qwen exhibits a starker version: the base model *degrades* under both ENT and ENT+SR, indicating that the prefix selection heuristic actively harms performance when diversity is high and the lightweight signal cannot compensate. The RLVR variant, by contrast, achieves the largest gain in the table at 63.78% under ENT+SR. The two families differ in overall responsiveness, with OLMo showing more uniform gains, but the underlying trend is shared: as alignment reduces diversity, depth-based refinement with lightweight signals becomes increasingly effective. For models where the strategy degrades performance, a stronger teacher signal for refinement may be necessary, an option we leave for future work.

Remark: We note that the low baseline for RLVR models is partly an artifact of response length: due to our fixed generation budget of 7168 tokens (max context length of 8192), a valid answer is extracted much less often compared to the base/instruct variants, which on average need fewer tokens. This also explains why exploration is dramatically effective: the refinement critique can signal that no answer was found, redirecting towards shorter, more conclusive derivations.

4 Discussion

We have argued that the generating model's diversity profile plays a vital role in determining the optimal exploration strategy. High-diversity base models require breadth; low-diversity aligned models benefit from depth. This pattern is consistent with published work: TreeLLM* (Hou et al., 2025) and Chain-in-Tree (Li, 2025) apply MCTS with lightweight refinement signals to instruct models, where our framework predicts depth is effective. In contrast, breadth-first approaches tend to utilize base models (Zhuang et al., 2025; Zhao et al., 2026), which exhibit much larger pass@k–pass@1 gaps, meaning that even mild improvements over the parallel sampling baseline can yield absolute gains. When MCTS does target stronger improvements, it typically relies on trained process reward models (Zhang et al., 2024; Uesato et al., 2022) that provide dense, per-step feedback, precisely the strong external signal our framework predicts is necessary for high-diversity models.

Alternate Views. We advocate for explicitly characterizing the model's diversity regime before selecting an exploration strategy. Our experiments validate this via inference on a fixed checkpoint, but during training the regime may shift as the policy evolves. Even if one does this characterization adaptively across training phases, note that repeated regime estimation is computationally expensive, and the gains from adaptation may not even justify this cost. Beyond diversity, other factors such as dataset composition, prompt design, and model scale, may interact with or confound the effects we attribute to the diversity regime. Even in our own experiments, we observed that prompt variation substantially altered the effectiveness of MCTS. Disentangling these remains an open question.

LIMITATIONS

Our framework assumes discrete latent approaches, simplifying a continuous trajectory space. The prefix-based assumption that early tokens commit to an approach may not hold universally. Our experiments cover mathematical reasoning on one benchmark; diversity profiles may differ for other domains. The feedback model treats refinement as constant per iteration, while real refinement likely has diminishing returns.

ACKNOWLEDGMENTS

This research used both the DeltaAI advanced computing and data resource, which is supported by the National Science Foundation (award OAC 2320345) and the State of Illinois, and the Delta advanced computing and data resource which is supported by the National Science Foundation (award OAC 2005572) and the State of Illinois.. Delta and DeltaAI are joint efforts of the University of Illinois Urbana-Champaign and its National Center for Supercomputing Applications.

REFERENCES

Yavuz Bakman, Sungmin Kang, Zhiqi Huang, Duygu Nur Yaldiz, Catarina G. Belém, Chenyang Zhu, Anoop Kumar, Alfy Samuel, Salman Avestimehr, Daben Liu, and Sai Praneeth Karimireddy. Uncertainty as feature gaps: Epistemic uncertainty quantification of llms in contextual question-answering, 2025. URL https://arxiv.org/abs/2510.02671.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, and et al. Lu. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning, 2025. URL https://arxiv.org/abs/2504.11456.

Zhenyu Hou, Ziniu Hu, Yujiang Li, Rui Lu, Jie Tang, and Yuxiao Dong. Treerl: Llm reinforcement learning with on-policy tree search, 2025. URL https://arxiv.org/abs/2506.11902.

Xinzhe Li. Chain-in-tree: Back to sequential reasoning in llm tree search, 2025. URL https://arxiv.org/abs/2509.25835.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023. URL https://arxiv.org/abs/2303.17651.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe

Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback, 2022. URL https://arxiv.org/abs/2211.14275.

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, 2023. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Sean Welleck, Amanda Bertsch, Matthew Finlayson, Hailey Schoelkopf, Alex Xie, Graham Neubig, Ilia Kulikov, and Zaid Harchaoui. From decoding to meta-generation: Inference-time algorithms for large language models, 2024. URL https://arxiv.org/abs/2406.16838.

Yixuan Even Xu, Yash Savani, Fei Fang, and J. Zico Kolter. Not all rollouts are useful: Down-sampling rollouts in llm reinforcement learning, 2025. URL https://arxiv.org/abs/2504.13818.

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

Jiarui Yao, Yifan Hao, Hanning Zhang, Hanze Dong, Wei Xiong, Nan Jiang, and Tong Zhang. Optimizing chain-of-thought reasoners via gradient variance minimization in rejection sampling and rl, 2025. URL https://arxiv.org/abs/2505.02391.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search, 2024. URL https://arxiv.org/abs/2406.03816.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards, 2026. URL https://arxiv.org/abs/2505.19590.

Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning, 2023. URL https://arxiv.org/abs/2305.14078.

Haomin Zhuang, Yujun Zhou, Taicheng Guo, Yue Huang, Fangxu Liu, Kai Song, and Xiangliang Zhang. Exploring multi-temperature strategies for token- and rollout-level control in rlvr, 2025. URL https://arxiv.org/abs/2510.08892.

Maciej Świechowski, Konrad Godlewski, Bartosz Sawicki, and Jacek Mańdziuk. Monte carlo tree search: a review of recent modifications and applications. *Artificial Intelligence Review*, 56(3): 2497–2562, July 2022. ISSN 1573-7462. doi: 10.1007/s10462-022-10228-y. URL http://dx.doi.org/10.1007/s10462-022-10228-y.

A NOTATION

Symbol	Meaning
Prefix-selection setting	
N, K, L	Total rollout budget, number of selected rollouts, and prefix length.
$x_{1:L}^{(i)}$	Length-L prefix of rollout i.
$\ell_t^{(i)}$, $m_t^{(i)}$	Chosen-token log-probability and top-1/top-2 log-probability gap at position t.
$q(i)$, $\hat{q}(i)$	Prefix score and its normalized version.
U_t, $\mathrm{disagree}(t)$	Number of distinct tokens and disagreement score at position t.
n, $\mathcal{T}$	Number and set of anchor positions.
S_i	Token set of prefix $x_{1:L}^{(i)}$.
$\mathrm{sim}_{\mathrm{Jac}}(i,j)$, $\mathrm{dist}_{\mathrm{Jac}}(i,j)$	Jaccard similarity and distance between prefixes i and j.
λ, α	Quality-diversity tradeoff and mixing weight.
$d(i,j)$, $d_{\mathrm{broad}}(i,j)$, $d_{\mathrm{deep}}(i,j)$	Hybrid, broad, and deep distances.
$\mathbf{p}_t^{(i)}$	Probability vector at anchor position t for rollout i.
$\mathrm{JSD}(\cdot,\cdot)$, $D_{\mathrm{KL}}(\cdot\|\cdot)$	Jensen–Shannon and Kullback–Leibler divergences.
$w_t(i,j)$	Anchor weight at position t.
$\omega(s)$	Step-dependent quality weight.
Proof notation for Proposition 1	
$\tau_1, \ldots, \tau_N$	Sampled rollouts/trajectories.
$p_\phi(\cdot \mid q)$	Model distribution conditioned on query q.
$\bar{\theta}$	Marginal success probability of one sampled rollout.
$\mathcal{A}_i$, π_i, θ_i	Approach class, its probability, and its success probability.
α, N_e, N_f	Exploration fraction, exploration budget, and refinement budget.
$i^\star$, $\theta_{i^\star}^{(t)}$	Best discovered approach and its stage-t success probability.
c_{rand}, $c_{i^\star}^{(t)}$, $R^{(t)}$	Random-sampling exponent, refinement exponent, and relative improvement ratio.
η, $\alpha_{\min}(\eta)$	Failure tolerance and minimum exploration fraction for discovery.
$\mathrm{pass@}N_{\mathrm{rand}}$, $\mathrm{pass@}N_{\mathrm{mcts}}$	Success probabilities of random sampling and the MCTS-style strategy.

B PREFIX SELECTION METHODS

Setting. For each problem, we sample N independent rollouts from the model and retain only the first L tokens (the *prefix*) of each rollout. A selection method receives only these N prefixes and must choose a subset of size $K \ll N$. Methods never use ground-truth correctness; labels are used only to evaluate $\mathrm{pass@}K$ (whether at least one selected rollout is correct).

Prefix features. For rollout $i \in [N]$, let $x_{1:L}^{(i)}$ denote its prefix tokens. We use: (i) chosen-token log-probabilities $\ell_t^{(i)} := \log p(x_t^{(i)} \mid x_{<t}^{(i)})$, (ii) top-$k$ logprob vectors at each position (when available), and (iii) the margin $m_t^{(i)} := \log p(\text{top-1 at } t) - \log p(\text{top-2 at } t)$.

Quality score. We define the prefix sequence log-probability and its normalised form

$$q(i) := \sum_{t=1}^{L} \ell_t^{(i)}, \qquad \hat{q}(i) := \frac{q(i) - \min_j q(j)}{\max_j q(j) - \min_j q(j)} \in [0,1]. \tag{1}$$

We use $\hat{q}(i)$ when mixing quality with diversity.

Token-disagreement anchors. Let $U_t := \left|\{x_t^{(i)} : i \in [N]\}\right|$ be the number of distinct tokens observed at position t across the N prefixes. Define

$$\mathrm{disagree}(t) := \frac{U_t - 1}{N - 1} \in [0,1]. \tag{2}$$

Given an anchor budget n, we select the n positions with the highest $\mathrm{disagree}(t)$ and denote the set by $\mathcal{T}$.

Jaccard similarity on prefix tokens. For rollout i, let $S_i := \{x_1^{(i)}, \ldots, x_L^{(i)}\}$ be the set of tokens appearing in its prefix. Define

$$\mathrm{sim}_{\mathrm{Jac}}(i,j) := \frac{|S_i \cap S_j|}{|S_i \cup S_j|}, \qquad \mathrm{dist}_{\mathrm{Jac}}(i,j) := 1 - \mathrm{sim}_{\mathrm{Jac}}(i,j). \tag{3}$$

B.1 RANDOM@K (BASELINE)

Method. Select K rollouts uniformly at random from the N samples. This provides the null baseline for whether prefix information is useful.

B.2 MMR (MAXIMAL MARGINAL RELEVANCE)

MMR greedily balances selecting *high-quality* prefixes while avoiding redundancy with already-selected prefixes.

Greedy rule. Let S be the selected set (initially empty) and R the remaining indices. At each step, choose

$$i^* \in \arg\max_{i \in R} \Big[\lambda\, \hat{q}(i) \;-\; (1-\lambda)\, \max_{j \in S} \mathrm{sim}_{\mathrm{Jac}}(i,j)\Big], \tag{4}$$

add i^* to S, and repeat until $|S| = K$. (When S is empty, the similarity term is taken as 0.)

Key hyperparameter: λ (quality–diversity tradeoff). The parameter $\lambda \in [0,1]$ controls how strongly the method prioritises high-probability prefixes versus novelty:

- $\lambda \approx 1$: *quality-dominant*. The method approaches selecting the top-K prefixes by $\hat{q}(i)$, using diversity only as a weak tie-breaker.
- $\lambda \approx 0$: *diversity-dominant*. The method prioritises reducing redundancy, even if it must accept lower-probability prefixes.

In our sweeps, smaller λ typically helps in high-diversity regimes (where many rollouts follow the same incorrect approach), while larger λ can help in low-diversity regimes (where the dominant approach is more often viable).

B.3 ADAPTIVE DISPERSION (TOKEN-DISAGREEMENT ANCHORS)

Adaptive Dispersion is a logprob-aware diversity selection method designed to distinguish *true approach forks* from superficial variation. It has three components: (1) find anchor positions where rollouts actually diverge, (2) define a hybrid distance that combines distributional differences at anchors with broad token-level differences, and (3) greedily select a set that transitions from quality to diversity as it fills the K slots.

B.3.1 STEP 1: CHOOSE ANCHOR POSITIONS

Compute disagreement scores (2) and select the top n positions:

$$\mathcal{T} := \text{top-}n \text{ positions by } \mathrm{disagree}(t). \tag{5}$$

Key hyperparameter: n (number of anchors). The integer n controls how many positions are treated as decision points.

- Small n: anchors focus on the most prominent early forks; this is robust but may miss secondary divergences.
- Large n: anchors capture finer-grained branching structure but can include noisy positions, which makes distance estimates less stable.

A useful rule of thumb is that n should be small relative to L and also not so large that most prefixes become unique "by chance" at the anchor positions.

B.3.2 STEP 2: HYBRID DISTANCE

Define a distance between prefixes i and j as

$$d(i,j) := \alpha\, d_{\text{deep}}(i,j) + (1-\alpha)\, d_{\text{broad}}(i,j), \qquad \alpha \in [0,1]. \tag{6}$$

Broad distance. We use Jaccard distance:

$$d_{\text{broad}}(i,j) := \text{dist}_{\text{Jac}}(i,j). \tag{7}$$

Deep distance (distributional at anchors). At each anchor position $t \in \mathcal{T}$, let $\mathbf{p}_t^{(i)}$ be the probability vector obtained by applying softmax to the stored top-k logprob vector for rollout i at position t (renormalised over the top-k support). We measure distributional disagreement via Jensen–Shannon divergence:

$$\text{JSD}(\mathbf{p},\mathbf{r}) := \frac{1}{2} D_{\text{KL}}(\mathbf{p} \,\|\, \mathbf{m}) + \frac{1}{2} D_{\text{KL}}(\mathbf{r} \,\|\, \mathbf{m}), \qquad \mathbf{m} = \frac{\mathbf{p}+\mathbf{r}}{2}. \tag{8}$$

We weight anchors by inverse margin:

$$w_t(i,j) := \frac{1}{1 + \frac{1}{2}\left(m_t^{(i)} + m_t^{(j)}\right)}. \tag{9}$$

Then

$$d_{\text{deep}}(i,j) := \frac{\sum_{t\in\mathcal{T}} w_t(i,j)\, \text{JSD}\left(\mathbf{p}_t^{(i)}, \mathbf{p}_t^{(j)}\right)}{\sum_{t\in\mathcal{T}} w_t(i,j)}. \tag{10}$$

Key hyperparameter: α (deep vs. broad distance). The mixing parameter $\alpha \in [0,1]$ trades off:

- $\alpha \approx 1$: distance is dominated by distributional differences at anchors (captures subtle "state-of-mind" differences even when sampled tokens match).
- $\alpha \approx 0$: distance reduces toward token-level Jaccard over the whole prefix (captures broad stylistic/lexical differences).

Empirically, larger α tends to help when anchor positions reliably correspond to genuine approach forks; smaller α can be more stable when token-level differences are spread across many positions.

B.3.3 STEP 3: GREEDY MAX–MIN SELECTION WITH AN ADAPTIVE SCHEDULE

Adaptive Dispersion builds S greedily using a max–min diversity term together with a step-dependent quality weight. Let $s \in \{0,1,\ldots,K-1\}$ be the selection step. Define a linear schedule

$$\omega(s) := \omega_{\text{init}}\left(1 - \frac{s}{K-1}\right) + \omega_{\text{final}}\frac{s}{K-1}, \qquad 0 \le \omega_{\text{final}} \le \omega_{\text{init}} \le 1. \tag{11}$$

At step s, select

$$i^* \in \arg\max_{i\in R} \Big[\omega(s)\, \hat{q}(i) + \big(1-\omega(s)\big)\, \min_{j\in S} d(i,j)\Big], \tag{12}$$

add i^* to S, and repeat. When S is empty, the diversity term is omitted and we pick the highest-quality prefix.

Key hyperparameters: $\omega_{\text{init}}, \omega_{\text{final}}$ (quality schedule). The schedule in (11) controls how quickly the method transitions from *exploitation* to *diversification*:

- Larger ω_{init}: the first few picks prioritise a strong "anchor" prefix (high $\hat{q}$).
- Smaller ω_{final}: later picks become close to pure diversity via the max–min term $\min_{j\in S} d(i,j)$.

A wide gap $\omega_{\text{init}} - \omega_{\text{final}}$ yields a stronger shift toward diversity as the set fills; a narrow gap keeps the method relatively quality-focused throughout.

Interaction with K. Because $\omega(s)$ depends on K, the same $(\omega_{\text{init}}, \omega_{\text{final}})$ can behave differently for different K values. For small K, the schedule has fewer steps and thus less opportunity to "cool down" from quality to diversity; for larger K, the later selections are more purely diversity-driven.

B.4 Method taxonomy (summary table)

Table 2 summarises the methods used in the main text by the signals they rely on.

Table 2: Taxonomy of prefix-selection methods used in the main text. "Top-k" indicates whether the method requires per-position top-k logprob vectors (beyond chosen-token logprobs).

Method	Quality signal	Diversity signal	Uses top-k?
Random@K	none	none	No
MMR	prefix logprob $\hat{q}(i)$	Jaccard over prefix tokens	No
Adaptive Dispersion	prefix logprob $\hat{q}(i)$	token-disagreement anchors + hybrid distance	Yes

B.5 Hyperparameter sweeps (values and reporting)

We sweep small grids and report the best-performing variant per model and K in Table 1 of the main text.

MMR. We sweep λ over a small set (e.g., $\{0.3, 0.5, 0.7\}$). Lower values emphasise novelty, while higher values emphasise selecting high-logprob prefixes first.

Adaptive Dispersion. We sweep:

- n (anchors): a small set such as $\{6, 8, 10\}$,
- α (deep weight): a small set such as $\{0.3, 0.5, 0.7\}$,
- $(\omega_{\text{init}}, \omega_{\text{final}})$ (schedule endpoints): a small set such as $(0.7, 0.1)$ and $(0.5, 0.1)$.

Across these variants, the method spans from *quality-first* (large ω_{init}) to *diversity-first* (small ω_{init}), and from *anchor-heavy* (large α) to *token-overlap-heavy* (small α).

C Proof of Proposition 1

Proof. We prove the ratio condition and the feasibility statement.

Under i.i.d. sampling, $\tau_1, \ldots, \tau_N \overset{\text{i.i.d.}}{\sim} p_\phi(\cdot \mid q)$ and each draw succeeds with marginal probability $\bar{\theta} := \Pr(\text{success})$. Hence

$$\text{pass@}N_{\text{rand}} = 1 - (1-\bar{\theta})^N = 1 - \exp\big(-Nc_{\text{rand}}\big), \qquad c_{\text{rand}} := -\log(1-\bar{\theta}).$$

If we further partition trajectory space into approaches $\{\mathcal{A}_i\}_{i=1}^m$ and define $\pi_i := \Pr(\tau \in \mathcal{A}_i)$ and $\theta_i := \Pr(\text{success} \mid \tau \in \mathcal{A}_i)$, then by the law of total probability $\bar{\theta} = \sum_{i=1}^m \pi_i \theta_i$.

Let $\alpha := N_e/N$ so $N_f = (1-\alpha)N$. Under the oracle identification assumption, exploitation is restricted to i^* and consists of N_f independent attempts, each succeeding with probability $\theta_{i^*}^{(t)}$. Therefore,

$$\text{pass@}N_{\text{mcts}} = 1 - (1-\theta_{i^*}^{(t)})^{N_f} = 1 - \exp\big(-N_f c_{i^*}^{(t)}\big), \qquad c_{i^*}^{(t)} := -\log(1-\theta_{i^*}^{(t)}).$$

Thus $\text{pass@}N_{\text{mcts}} \geq \text{pass@}N_{\text{rand}}$ is equivalent to

$$\exp\big(-(1-\alpha)Nc_{i^*}^{(t)}\big) \leq \exp\big(-Nc_{\text{rand}}\big) \iff (1-\alpha)c_{i^*}^{(t)} \geq c_{\text{rand}} \iff \alpha \leq 1 - \frac{c_{\text{rand}}}{c_{i^*}^{(t)}}.$$

With $R^{(t)} := c_{i^*}^{(t)}/c_{\text{rand}}$, this becomes $\alpha \leq 1 - 1/R^{(t)}$.

During exploration, each of the $N_e = \alpha N$ samples lands in approach i^* with probability π_{i^*}, so

$$P(\text{discover } i^*) = 1 - (1 - \pi_{i^*})^{N_e} = 1 - (1 - \pi_{i^*})^{\alpha N}.$$

To ensure $P(\text{discover } i^*) \geq 1 - \eta$, it suffices that $(1 - \pi_{i^*})^{\alpha N} \leq \eta$, i.e.,

$$\alpha \geq \frac{\log(1/\eta)}{N \log\left(\frac{1}{1-\pi_{i^*}}\right)} =: \alpha_{\min}(\eta),$$

and using $\log\left(\frac{1}{1-x}\right) \approx x$ for small x gives $\alpha_{\min}(\eta) \approx \frac{\log(1/\eta)}{N\pi_{i^*}}$. Finally, an α satisfying both discovery and outperformance exists iff $\alpha_{\min}(\eta) \leq 1 - \frac{1}{R^{(t)}}$, equivalently $R^{(t)} \geq \frac{1}{1-\alpha_{\min}(\eta)}$. □