ASSESSING DIVERSITY COLLAPSE IN REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

We identify a striking phenomenon in large language models finetuned on reasoning tasks: as Pass@1 improves during supervised finetuning, Pass@k rapidly deteriorates and fails to recover with reinforcement learning or self-improvement. We formalize the relationship between expected Pass@k and Pass@1 over the test distribution and attribute the early drop in Pass@k to diversity collapse—where fine-tuning causes the probability mass to converge toward a single reasoning path and final answer for test questions. We theoretically prove how the standard finetuning strategy of SFT and RL leads to diversity collapse in reasoning models. Finally, we estimate the optimal Pass@k performance achievable with an oracle given access to the model's distribution over final answers marginalized over all rollouts and reveal a significant gap compared to current token-level diverse decoding methods such as temperature scale, top-k, nucleus, and min-p sampling. We highlight the need for better decoding strategies for generating reasoning steps during self-improvement and inference. Finally, we propose a promising solution by model weight interpolation.

1 INTRODUCTION

Recent advances in large language models (LLMs) have showcased their remarkable ability to perform complex mathematical reasoning (Lightman et al., 2023; Azerbayev et al., 2023), yet these successes often hinge on test-time scaling strategies(Snell et al., 2024; Wu et al., 2024). In many applications, particularly those involving math problems, puzzles, and logical reasoning, LLMs employ a verifier-generator framework in which it is significantly easier for the model to verify a candidate solution than to generate one from scratch. This distinction has given rise to strategies that leverage multiple rollouts during inference, selecting the best answer through an outcome reward model (ORM) (Lightman et al., 2023). Consequently, performance is typically measured using the Pass@k metric—the probability that at least one out of k sampled answers is correct—rather than solely relying on Pass@1. (Li et al., 2024) This shift in evaluation underscores the need for robust decoding strategies that enhance diversity in the generated solutions while maintaining high accuracy. Despite the impressive gains observed with test-time scaling, current training and decoding paradigms are suboptimal with respect to maximizing Pass@k. Empirical evidence suggests that while techniques such as Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) can incrementally improve the quality of the best single rollout (Pass@1), they often lead to a degradation in overall diversity, causing Pass@k to drop. (Cobbe et al., 2021)

In this paper, we rigorously analyze the gap between expected Pass@1 and Pass@k over the test distribution. First, we formalize how Pass@k depends on both the mean and variance of per-example Pass@1 probabilities. We highlight an inherent asymmetry in Pass@k where a drop in Pass@1 for any given question is disproportionately detrimental compared to an equivalent gain for another. Our experiments confirm prior observations (e.g., on GSM8K (Cobbe et al., 2021)) where Pass@1 improvements don't reliably translate to Pass@k gains due to this variance asymmetry. Furthermore, we empirically and theoretically establish that this variance in Pass@1 is caused by SFT and RL tending to collapse the diversity of the model's generations to *a single reasoning trace*: SFT miscalibrates models toward oversampling a common response, while RL fails to restore diversity. Finally, we estimate the optimal Pass@k achievable with an oracle given access to the model's distribution over final answers marginalized over all rollouts and reveal a significant gap of $10 - 20\%$ compared to current token-level diverse decoding methods like temperature scale, top-k(Shao et al., 2017), nucleus(Holtzman et al., 2020), and min-p(Nguyen et al., 2024) sampling. Furthermore, we propose a method to mitigate diversity collapse during SFT by interpolating the model weights from earlier and later checkpoints, which demonstrates superior performance in Pass@k accuracy.

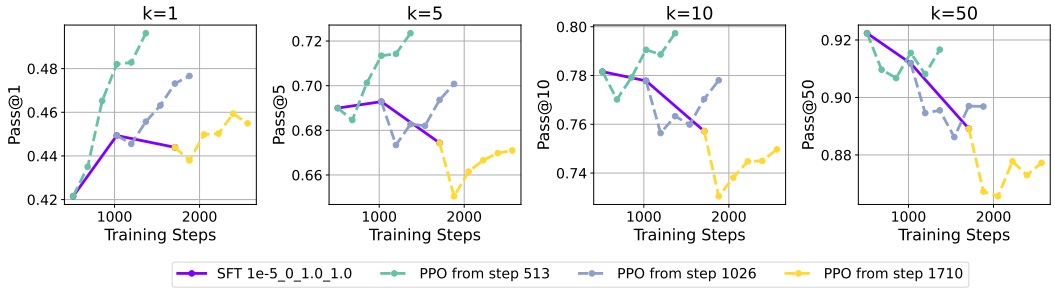

Figure 1: **Pass@k for SFT and RL checkpoints of Qwen-2.5-0.5B on GSM8K.** The purple solid line is measures the performance across SFT steps, while the three dashed lines correspond to RL with Proximal Policy Optimization (PPO) starting from different SFT checkpoints. While Pass@1 continues to improve, Pass@k for larger k oftentimes decreases, even with RL.

2 Asymmetry of Pass@k

We first formalize the relationship between Pass@k and Pass@1, and why these metrics can be anti-correlated. Given a reasoning model $f(\cdot)$, a decoding strategy D, and a question x, the model's answer $\hat{y}$ is obtained by sampling a *rollout* $\boldsymbol{r} := [x, s^{(1)}, s^{(2)}, ..., s^{(n)}, \hat{y}]$ where $s^{(i)}$ are intermediate steps. Then given k rollouts, Pass@k measures if one of the guesses is equal to the true answer y. If these rollouts are sampled i.i.d., the expected *per-example* Pass@k is equal to

$$\text{Pass@k}(x) = \mathbb{E}_{[\boldsymbol{r}_i]_{i=1}^{k} \sim D(f(x))} \left[\mathbb{1}\{\exists\, i \text{ s.t. } \hat{y} = y\}\right] = 1 - (1 - \rho_x)^k \tag{1}$$

where $\rho_x = P(\hat{y} = y \mid x, f, D)$ is the marginal probability of sampling the ground truth answer or what we call the "ground truth confidence".

We now provide a straightforward upper bound of *expected* Pass@k over the entire test distribution $x, y \sim \mathcal{D}$, that depends on the *expectation and variance* of Pass@1 (Proof in Appendix B).

Proposition 2.1. $\mathbb{E}_{x,y\sim\mathcal{D}}\left[\text{Pass@k}(x)\right] \leq 1 - ((1 - \mathbb{E}_{x,y\sim\mathcal{D}}[\rho_x])^2 + \text{Var}(\rho_x))^{k/2}$

Notably, the upper bound of Pass@k rises with the expected Pass@1 but falls with its variance. As we will observe, finetuning reasoning models by SFT tends to increase *both* the expectation and variance of Pass@1, causing Pass@k to oftentimes decrease as Pass@1 increases.

3 Early Drop of Pass@k during SFT-RL

In Figure 1, we plot test Pass@k (k = 1, 5, 10, 50) of Gemma-2-2B and Qwen-2.5-0.5B trained on the rephrased augmentations of GSM8k (Cobbe et al., 2021) in MetaMathQA (**?**). A clear pattern emerges across both models and datasets: during SFT, Pass@1 continues to improve, but Pass@k for larger k peaks and then drop sharply. We then continue finetuning SFT checkpoints using Proximal Policy Optimization (PPO) with a binary reward of the correctness of the model's final answer. Although PPO successfully recovers Pass@k for early SFT checkpoints, the performance of later checkpoints often declines even further, especially for large k.

Bimodal ground truth confidence Why does Pass@k drop? We estimate the ground truth confidence ρ_x of the GSM8k test examples empirically over 100 rollouts sampled with temperature set to 1. In Figure 2, we plot the distribution of ground truth confidences and observe that the distribution becomes very *bimodal*. Consequently, the variance of Pass@1 increases. This increase in variance directly explains the drop in Pass@k as we saw from our upper bound in Proposition 2.1.

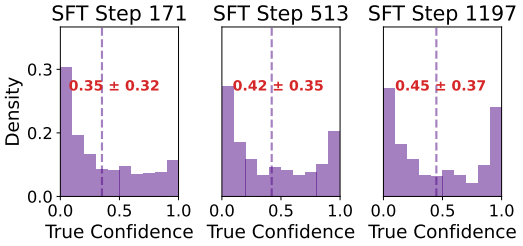

Figure 2: **Histogram of ρ_x** of Qwen-2.5-0.5B SFT checkpoints across GSM8k test.

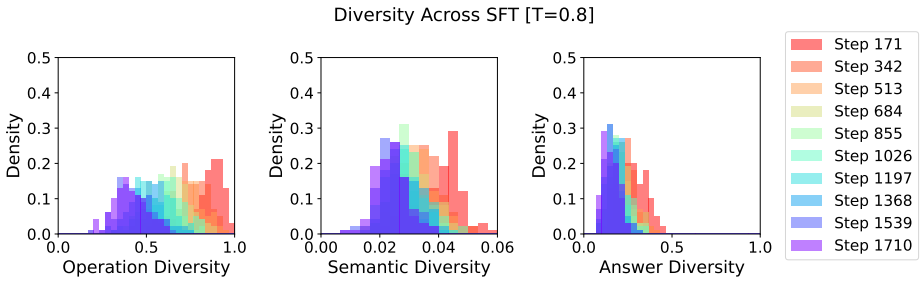

Figure 3: **Diversity Collapse** The answer, semantic, and operation diversity of the rollouts of Gemma-2-2B checkpoints across GSM8k test examples. Colors map to different SFT steps. As SFT progresses, diversity of intermediate steps and final answers collapses.

3.1 Diversity Collapse Towards a Single Reasoning Trace

The bimodal nature of the distribution indicates that, for each test example, the model's ground truth confidence is either very high or very low. One possible explanation for this behavior is if the model is highly confident about its predictions: if it is confidently wrong, the ground truth confidence is low, whereas if it is confidently correct, the ground truth confidence is high.

To test this hypothesis, we sample 100 rollouts per GSM8k test example from Gemma-2-2B and measure the degree of *diversity collapse* between the example's rollouts using three metrics: 1.) **Answer Diversity:** the fraction of unique guesses $\hat{y}$ among rollouts, 2.) **Semantic Diversity:** the average cosine similarity between the text embeddings of the rollouts, computed using Stella-400M-v5 (Zhang et al., 2024a), 3.) **Operation Diversity:** group rollouts by the sequence of arithmetic operations performed and measure the fraction of unique operation sequences.

In Figure 3, we observe that longer SFT leads to a clear collapse in the model's probability distribution. The model places most of its probability mass on one answer and *a single reasoning trace*, as evidenced by reduced semantic and operation diversity.

4 Theoretical Limitations of SFT and RL

Diversity Collapse in Supervised Finetuning Overparameterized models are well-known to exhibit overconfidence in their predictions, an effect that has been studied extensively in classification (Guo et al., 2017). In particular, the model's confidence towards the most likely class $P(\hat{y} = k_{\max} \mid x)$ is often much higher than the model's accuracy. In binary classification with linear models $f(x) = \sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$ and linearly separable training data, gradient descent provably drives the norm of the weights to infinity, causing probabilities to collapse to 0 or 1 (Soudry et al., 2018). We plot this explicitly in Appendix Figure 6 in models trained on binary Gaussian mixture data. A similar phenomenon likely arises in large reasoning models, which are also finetuned on relatively small datasets by cross-entropy loss, ultimately leading to overly confident solutions in spite of limited coverage over the space of rollouts (Cobbe et al., 2021).

Diversity Collapse in Reinforcement Learning Our analysis shows that applying standard reinforcement learning algorithms, such as GRPO (Ramesh et al., 2024), over reasoning models with low initial diversity can inevitably collapses output diversity. We quantify initial diversity by the sharpness of the bimodal ground truth confidence (i.e., accuracy) distribution: when most samples yield either very low or very high ground truth confidence, the RL dynamics exhibit two notable phenomena: 1.) The probability mass on incorrect rollouts decays rapidly to negligible levels, 2.) The remaining probability mass over the correct rollouts evolves with approximately zero drift but finite variance.

These dynamics create a *winner-takes-all effect*. Once one correct reasoning trace is randomly reinforced, its probability begins to dominate while the gradient signal for other traces diminishes. Consequently, the maximum probability among the correct traces eventually exceeds a preset threshold $\Delta \gg \frac{1}{K}$ (where K denotes the number of potential correct reasoning traces). This implies that the model concentrates nearly all of its probability mass on a single correct reasoning trace for each sample in the training set, effectively collapsing the system into an SFT-like regime.

Theorem 4.1 (Diversity Collapse in Reinforcement Learning (Informal)). *Suppose we are training a model with GRPO with group size G on a finite set of M problems. For each problem, there are multiple correct traces. Assume the initial accuracy distribution of the samples is drawn from a sharp beta distribution with Beta$(a, b), a, b < \tau \ll 1$. We say the model is collapse when the probability of one correct reasoning path is dominant for each problem, and let $\mathcal{C}$ denote the event of collapse, then we have*

$$Pr(\mathcal{C}) \geq 1 - M\left(1 - \exp\left(-\frac{CG}{\eta^2 \mathcal{O}(\tau)}\right)\right) \tag{2}$$

where η is the learning rate and C is a constant.

5 Limitations of Token-Level Diverse Decoding Strategies

Furthermore, we analyze how well standard decoding strategies for diversity helps recover the Pass@k drop and compare against the *optimal* Pass@k strategy. Intuitively, to achieve high Pass@k, a model's optimal strategy is to sample k *unique and most likely* guesses. In the ideal scenario where the decoder had oracle access to the model's marginal distribution over final answers $P(\hat{y} = a \mid x)\ \forall a \in \mathcal{A}$, we could use these probabilities to either sample answers without replacement or sample top-K answers. This is often challenging in practice as autoregressive models do not have any foresight about the final answer distribution of any intermediate step.

However, we can estimate the optimal Pass@k of Gemma-2-2B and Qwen-2.5-0.5B in GSM8k by using the empirical answer distribution over 1000 sampled rollouts per example. We try decoding with a range of temperatures $T \in [0.8, 1.8]$. First, in Figures 10 and 13, we surprisingly notice a stark gap of $10 - 20\%$ between the best Pass@K$(K = 2, 4, 8)$ achieved by naive decoding with temperature scaling and the top-K w/oracle sampling. For example, the optimal Pass@2 of Gemma-2-2B is close to 84%. This suggests that current decoding strategies significantly underestimate a reasoning model's true performance. In Tables 1 and 2, we also compare standard diversity-inducing sampling strategies min-p (Nguyen et al., 2024), nucleus (Holtzman et al., 2020), and top-k(Shao et al., 2017). We find that these methods do not do significantly better than carefully tuning the temperature. Token-level diversity strategies tend to have a strict trade-off between either sampling similar rollouts with the same answer or increasing the likelihood of sampling outlier answers.

6 Weight Interpolation

We propose a method that could help fix the diversity issue of LLM through weight interpolation between earlier checkpoints and later checkpoints. Specifically, we run a uniform weight averaging (Wortsman et al., 2022) of each checkpoint with an early checkpoint at SFT Step 171. In Figure 16, we see that decoding from a weight-interpolated model is superior to baseline decoding strategies for Pass@k. In particular, there is no drop in the curve of pass@k during SFT. This suggests that weight interpolation can minimize the variance of the confidence distribution in 2.1 without significantly decreasing the expectation (i.e, Pass@1) like temperature scaling. By "mixing" these confidence distributions, the diverisity of sampled rollouts significantly improves without oversampling outliers.

7 Conclusion

In conclusion, our findings shed light on the trade-off between optimizing for Pass@1 and maintaining reasoning diversity, emphasizing that conventional finetuning and reinforcement learning strategies inadvertently lead to diversity collapse. We also highlight the limitations of token-level decoding strategies for generating reasoning steps during self-improvement and inference. For future work, we believe it may be important to further investigate model weight interpolation and its efficacy for improving Pass@k compare to other diverse decoding strategies specifically designed for reasoning. This could hold the potential to bridge the gap between current decoding limitations and the optimal oracle performance that could be extracted from current models.

REFERENCES

Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edition, 1964.

AlphaProof and AlphaGeometry teams. Ai achieves silver-medal standard solving international mathematical olympiad problems, jul 2024. URL https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/. Google DeepMind blog post.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*, 2023.

Edward Beeching, Lewis Tunstall, and Sasha Rush. Scaling test-time compute with open models. URL https://huggingface.co/spaces/HuggingFaceH4/blogpost-scaling-test-time-compute.

Jeff Bilmes. Submodularity in machine learning and artificial intelligence. *arXiv preprint arXiv:2202.00132*, 2022.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training, 2025. URL https://arxiv.org/abs/2501.17161.

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, 2021. URL https://arxiv.org/abs/2110.14168.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

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

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

I. Karatzas and S. Shreve. *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics (113) (Book 113). Springer New York, 1991. ISBN 9780387976556.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making large language models better reasoners with step-aware verifier, 2023. URL https://arxiv.org/abs/2206.02336.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26763–26773, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

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, 2025. URL https://arxiv.org/abs/2501.19393.

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, 2024. URL https://arxiv.org/abs/2407.01082.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free rlhf, 2024. URL https://arxiv.org/abs/2405.20304.

Alex Renda, Aspen Hopkins, and Michael Carbin. Can llms generate random numbers? evaluating llm sampling in controlled domains. Technical report, MIT CSAIL, Cambridge, MA, 2023. URL https://people.csail.mit.edu/renda/llm-sampling-paper. ICML 2023 Workshop: Sampling and Optimization in Discrete Space (SODS 2023). Available at: https://people.csail.mit.edu/renda/llm-sampling-paper.

Mikayel Samvelyan, Sharath Chandra Raparthy, Andrei Lupu, Eric Hambro, Aram H. Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, Tim Rocktäschel, and Roberta Raileanu. Rainbow teaming: Open-ended generation of diverse adversarial prompts, 2024. URL https://arxiv.org/abs/2402.16822.

Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot, Johan Ferret, Nino Vieillard, Alexandre Ramé, Bobak Shariari, Sarah Perrin, Abe Friesen, Geoffrey Cideron, Sertan Girgin, Piotr Stanczyk, Andrea Michi, Danila Sinopalnikov, Sabela Ramos, Amélie Héliou, Aliaksei Severyn, Matt Hoffman, Nikola Momchev, and Olivier Bachem. Bond: Aligning llms with best-of-n distillation, 2024. URL https://arxiv.org/abs/2407.14622.

Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. Generating high-quality and informative conversation responses with sequence-to-sequence models. *arXiv preprint arXiv:1701.03185*, 2017.

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.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. PMLR, 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.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms, 2025. URL https://arxiv.org/abs/2502.03373.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. Jasper and stella: distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*, 2024a.

Yiming Zhang, Avi Schwarzschild, Nicholas Carlini, Zico Kolter, and Daphne Ippolito. Forcing diffuse distributions out of language models, 2024b. URL https://arxiv.org/abs/2404.10859.

A Literature Review

We review recent works on (1) evaluation via Pass@k metrics and diverse test-time strategies and (2) fine-tuning pipelines for reasoning that combine supervised fine-tuning (SFT) with reinforcement learning (RL).

A.1 Pass@k Evaluation and Test-Time Strategies

Recent studies have shown that allocating additional compute at test time can be as effective as scaling model size. Snell et al. (2024) demonstrated that increasing the test-time compute budget—by allowing more intermediate reasoning steps, employing majority voting, or using process reward models—can substantially boost model performance. In practice, Pass@N is used as an upper bound metric for the best-of-N performance, where a verifier, often superior to the generator, selects the best candidate among several rollouts (Cobbe et al., 2021).

Recently, several works have proposed methods to optimize best-of-N strategies. Sessa et al. (2024) proposes a Best-of-N-aware finetuning strategy that explicitly trains models to output diverse rollouts. Similarly, Muennighoff et al. (2025) found that linear search encourages diversity and tends to perform much better than parallel decoding. To improve diversity during parallel decoding, Beeching et al. proposes a diverse beam search decoding strategy and Li et al. (2023) appends curated prompts to the generation. In some cases, when models are trained to work with formal languages (e.g., Lean), methods such as Monte Carlo tree search can explicitly diversify reasoning steps in rollouts, such as in AlphaProof (AlphaProof and AlphaGeometry teams, 2024).

Outside the direct reasoning framework, efforts to achieve a more diffuse output distribution include methods based on diverse fine-tuning (Zhang et al., 2024b) and decoding strategies. Techniques such as min-p sampling (Nguyen et al., 2024) and nucleus sampling (Holtzman et al., 2020) have been studied to maintain realistic language outputs while still promoting creative diversity. Moreover, prompting strategies that append either random or manually curated prompts to the original question have been shown to enforce varied solutions (Samvelyan et al., 2024; Renda et al., 2023).

A.2 Fine-Tuning Pipelines for Reasoning

The conventional pipeline for enhancing reasoning in LLMs involves an initial phase of supervised fine-tuning (SFT) followed by reinforcement learning (RL). SFT is critical for instilling interpretable reasoning chains and ensuring that the model adheres to a consistent rollout template. However, recent work has raised concerns regarding SFT's potential pitfalls. Yeo et al. (2025) critically examine the necessity of SFT, showing that while it guides reasoning, it may also induce overfitting to the training distribution. This overfitting can limit a model's ability to generalize, as further evidenced by Chu et al. (2025), who report that SFT-trained models tend to memorize training data, resulting in higher out-of-distribution (OOD) failures compared to those refined via RL. Notably, Cobbe et al. (2021) similarly observe that Pass@k can drop before Pass@1 in GSM8k.

In parallel, the Deepseek-r1 approach proposed by Guo et al. (2025) leverages RL to directly incentivize reasoning capability, challenging the assumption that SFT is always beneficial. Despite these critiques, SFT continues to play a foundational role in many reasoning models, as it provides an interpretable and structured chain-of-thought that remains crucial during early training stages. In our work, we extend these findings by analyzing how early-stopping SFT based on Pass@1 validation metrics can lead to suboptimal performance on Pass@k, and establish a formal relationship between these evaluation metrics. We further compare current decoding strategies against an oracle Pass@k approach that utilizes full knowledge of the model's final answer distribution.

B EXPECTED PASS@K

Proposition B.1.

$$\mathbb{E}\left[\text{Pass@k}(\rho_x)\right] \leq 1 - \left((1 - \mathbb{E}[\rho_x])^2 + \text{Var}(\rho_x)\right)^{k/2}$$

Proof.

$$\mathbb{E}\left[(1-\rho_x)^k\right] \leq \mathbb{E}\left[(1-\rho_X)^2\right]^{k/2} \tag{3}$$

$$= (1 - 2\mathbb{E}\left[\rho_x\right] + \mathbb{E}\left[\rho_x^2\right])^{k/2} \tag{4}$$

$$= ((1 - 2\mathbb{E}\left[\rho_x\right] + \mathbb{E}\left[\rho_x\right]^2) + (\mathbb{E}\left[\rho_x^2\right] - \mathbb{E}\left[\rho_x\right]^2))^{k/2} \tag{5}$$

$$= \left((1 - \mathbb{E}\left[\rho_x\right])^2 + \mathsf{Var}(\rho_x)\right)^{k/2} \tag{6}$$

□

C SFT IN BINARY CLASSIFICATION

Data and Model Setup We train a linear classifier $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ from random initialization over a binary Gaussian mixture distribution:

$$x \mid y \sim \mathcal{N}(y\boldsymbol{\mu}, I^{d\times d}) \tag{7}$$

$$y \in \{1, -1\} \text{ uniformly} \tag{8}$$

Given a model, we sample predictions, namely $\hat{y} = 1$ with probability $\sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle) = (1 + \exp(-\langle \boldsymbol{w}, \boldsymbol{x} \rangle))^{-1}$, or $\hat{y} = 0$. Then, per-example Pass@1 is equal to $\rho_x = \sigma(y \cdot \langle \boldsymbol{w}, \boldsymbol{x} \rangle)$. Similarly, the expected Pass@k is equal to $1 - (1 - \rho_x)^k$.

In our experiment, we train an overparametrized linear classifier over binary Gaussian data mixture $x \mid y \sim \mathcal{N}(y \cdot \frac{1}{\sqrt{d}}\mathbf{1}, \frac{1}{2}I)$ where $y = \{-1, 1\}$ and $d = 1000$. We then evaluate ρ_x of 400 test samples. As training progresses, the distribution of ρ_x over the test data become bimodal due to the norm of w monotonically increasing once it separates the training examples. Similarly, we observe that this leads to a drop in Pass@k while Pass@1 continues to improve.

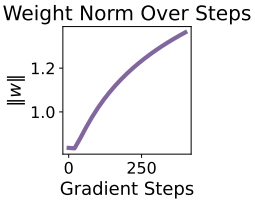

Figure 4: Weight Norm

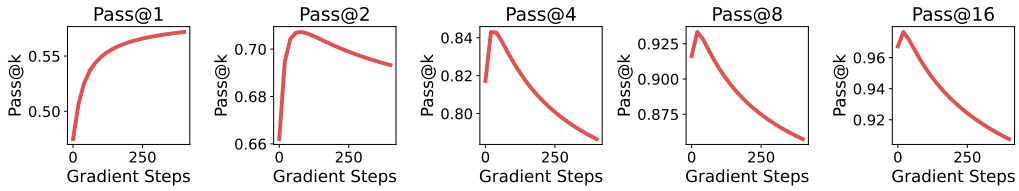

Figure 5: Pass@k across Training in Binary Classification

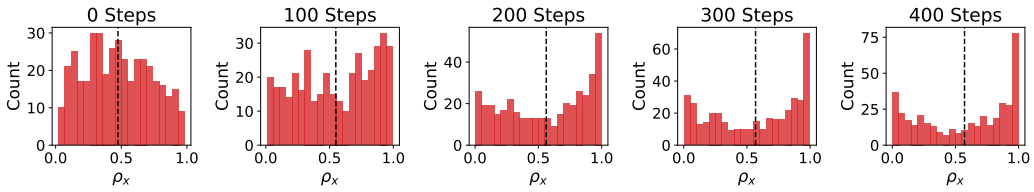

Figure 6: Histogram of ρ_x

D RL THEORY

Overview. We refine the analysis of how Proximal Policy Optimization (PPO) can collapse into an *SFT-like regime* when the policy starts with inadequate diversity. The core idea remains that initially *polarized* probabilities prevent large-scale exploration of new correct paths. However, we now incorporate *gradient interference* and *token-wise similarity* among paths in a neural network, showing that these effects can *reinforce* or *hinder* probability updates and ultimately sharpen the same collapse phenomenon.

D.1 DETAILED ANALYSIS OF DIVERSITY COLLAPSE IN GRPO

D.1.1 SETTING AND NOTATION

Consider a $(K+1)$-armed bandit with K "good" arms and 1 "bad" arm (labeled $K+1$). Let p_i be the probability of selecting arm i, so $\sum_{i=1}^{K+1} p_i = 1$. Denote the bad arm probability by

$$p_{K+1} \;=\; 1-\alpha,$$

where $\alpha = \sum_{i=1}^{K} p_i$ is the total probability of all good arms. Each iteration of training proceeds as follows:

1. Sample G arms $\{I_t^{(1)}, \ldots, I_t^{(G)}\}$ i.i.d. from the current policy $p(\cdot)$.
2. Observe rewards $r_t^{(g)} = \mathbf{1}\{I_t^{(g)} \le K\}$, where 1 indicates a good arm.
3. Compute

$$\mu_t \;=\; \frac{1}{G}\sum_{g=1}^{G} r_t^{(g)}, \quad \sigma_t \;=\; \sqrt{\frac{1}{G}\sum_{g=1}^{G}\big(r_t^{(g)}-\mu_t\big)^2},$$

   and define the normalized advantage

$$\widetilde{r}_t^{(g)} \;=\; \begin{cases} \dfrac{r_t^{(g)}-\mu_t}{\sigma_t}, & \sigma_t \neq 0, \\ 0, & \sigma_t = 0. \end{cases}$$

4. Update each parameter θ_i via

$$\theta_i \;\leftarrow\; \theta_i + \frac{\eta}{G}\sum_{g=1}^{G} \widetilde{r}_t^{(g)}\Big(\mathbf{1}\{I_t^{(g)} = i\} - p_i\Big).$$

5. Stop training when $p_{K+1} = 1-\alpha$ drops below a threshold $\beta > 0$.

We denote the stopping time by

$$T_{\text{stop}} \;=\; \inf\{\, t : p_{K+1}(t) < \beta\}.$$

We analyze the probability that some good arm probability p_i (with $1 \le i \le K$) exceeds a given threshold $\Delta > \frac{1}{K}$. Define the *collapse event*

$$\mathcal{C} \;=\; \left\{ \max_{1\le i\le K} p_i \;\ge\; \Delta \right\}.$$

In what follows, we derive both *upper* and *lower* bounds on $\Pr(\mathcal{C})$ under GRPO.

D.1.2 BAD ARM DRIFT AND STOPPING TIME

Lemma D.1 (Bad Arm Drift)**.** *Let m_t be the number of bad arms (arm $K+1$) sampled in the group at iteration t. Then the update increment of θ_{K+1} satisfies*

$$\Delta\theta_{K+1} \;=\; \frac{\eta}{G}\sum_{g=1}^{G} \widetilde{r}_t^{(g)}\Big(\mathbf{1}\{I_t^{(g)} = K+1\} - p_{K+1}\Big).$$

If $m_t \sim \text{Binomial}(G,\, 1-\alpha)$*, then*

$$\mathbb{E}\Big[\Delta\theta_{K+1}\Big] \;=\; \frac{\eta}{G}\,\mathbb{E}[X],$$

where

$$X \;=\; \sum_{g=1}^{G} \widetilde{r}_t^{(g)}\big(\mathbf{1}\{I_t^{(g)} = K+1\} - p_{K+1}\big).$$

Moreover, a more precise conditional analysis shows that

$$\mathbb{E}[X] \;\approx\; -G\,\sqrt{\alpha\,(1-\alpha)},$$

thus

$$\mathbb{E}\Big[\Delta\theta_{K+1}\Big] \;\approx\; -\,\eta\,\sqrt{\alpha\,(1-\alpha)}.$$

Proof. Condition on having m bad arms in the group, where $m \sim \text{Binomial}(G,\, 1-\alpha)$. One computes (since a "bad arm" has reward 0, and a "good arm" has reward 1, so the sample average reward is $\mu_t = \frac{G-m}{G}$, and sample variance is tied to m and $G-m$):

$$\widetilde{r}_{\text{bad}} \;=\; -\sqrt{\frac{G-m}{m}}, \quad \widetilde{r}_{\text{good}} \;=\; +\sqrt{\frac{m}{G-m}}.$$

For each bad arm sample, the factor $\mathbf{1}\{I^{(g)} = K+1\} - p_{K+1} = \alpha$, and for each good arm sample, the factor equals $-(1-\alpha)$. Summing yields

$$X = m\,\widetilde{r}_{\text{bad}}\,\alpha \;+\; (G-m)\,\widetilde{r}_{\text{good}}\,\big[-(1-\alpha)\big].$$

Substituting the expressions for $\widetilde{r}_{\text{bad}}$ and $\widetilde{r}_{\text{good}}$, one obtains

$$X = -\sqrt{m(G-m)}\big[\alpha + (1-\alpha)\big] = -\sqrt{m(G-m)}.$$

Hence

$$\mathbb{E}[X] \;=\; -\,\mathbb{E}\big[\sqrt{m(G-m)}\big].$$

Since $m \sim \text{Binomial}(G, 1-\alpha)$ has mean $G(1-\alpha)$ and variance $G\alpha(1-\alpha)$, a standard Delta-method or Taylor expansion around $m = G(1-\alpha)$ shows

$$\mathbb{E}\big[\sqrt{m(G-m)}\big] \;\approx\; G\,\sqrt{\alpha(1-\alpha)},$$

for large G. Consequently,

$$\mathbb{E}[X] \;\approx\; -\,G\,\sqrt{\alpha(1-\alpha)}.$$

Multiplying by $\frac{\eta}{G}$ yields

$$\mathbb{E}\big[\Delta\theta_{K+1}\big] \;\approx\; -\,\eta\,\sqrt{\alpha(1-\alpha)},$$

as claimed. □

Lemma D.2 (Bad Arm Probability Decay)**.** *Under the GRPO dynamics, the expected change in the bad arm probability satisfies:*

$$\mathbb{E}[\Delta p_{K+1}] \approx -\eta \cdot \alpha^{3/2}(1-\alpha)^{3/2},$$

where $\alpha = 1 - p_{K+1}$*. The continuous-time approximation leads to the differential equation:*

$$\frac{dp}{dt} = -\eta \cdot (1-p)^{3/2} p^{3/2}.$$

The exact stopping time to reach $p_{K+1} < \beta$ *is given by:*

$$T_{\text{stop}} = \frac{1}{\eta}\int_{\beta}^{1-\alpha} \frac{dp}{(1-p)^{3/2}p^{3/2}}.$$

Evaluating this integral yields:

$$T_{\text{stop}} = \frac{2}{\eta}\left[\frac{(2p-1)}{\sqrt{p(1-p)}}\right]_{p=\beta}^{p=1-\alpha}.$$

Expanding for general α *and* β*, this contains all* α*-dependent terms. For small* β *(*$\beta \to 0$*), the dominant term is:*

$$T_{\text{stop}} \sim \frac{2}{\eta\sqrt{\beta}}\,(1 + \mathcal{O}(\beta)) - \frac{2}{\eta\sqrt{\alpha(1-\alpha)}} + \mathcal{O}(1).$$

Thus, the stopping time retains explicit dependence on initial conditions α *through the second term.*

Proof. From Lemma D.1, the expected parameter update is:

$$\mathbb{E}[\Delta\theta_{K+1}] \approx -\eta\sqrt{\alpha(1-\alpha)}.$$

Using the softmax derivative relationship $\Delta p_{K+1} \approx p_{K+1}(1-p_{K+1})\Delta\theta_{K+1} = \alpha(1-\alpha)\Delta\theta_{K+1}$, we get:

$$\mathbb{E}[\Delta p_{K+1}] \approx -\eta \cdot \alpha^{3/2}(1-\alpha)^{3/2}.$$

Treating the discrete updates as a continuous process:

$$\frac{dp}{dt} = -\eta \cdot (1-p)^{3/2}p^{3/2}.$$

Separating variables and integrating:

$$\int_{1-\alpha}^{\beta} \frac{dp}{(1-p)^{3/2}p^{3/2}} = -\eta \int_0^{T_{\text{stop}}} dt.$$

Using the substitution $p = \sin^2\theta$:

$$\int \frac{dp}{(1-p)^{3/2}p^{3/2}} = 2\int \csc^2(2\theta)d\theta = -\frac{2(2p-1)}{\sqrt{p(1-p)}} + C.$$

Applying limits:

$$T_{\text{stop}} = \frac{2}{\eta}\left[\frac{(2p-1)}{\sqrt{p(1-p)}}\right]_{p=\beta}^{p=1-\alpha}.$$

For small β ($\beta \ll 1$):

$$\frac{(2\beta-1)}{\sqrt{\beta(1-\beta)}} \approx -\frac{1}{\sqrt{\beta}}\left(1 + \frac{3\beta}{2} + \cdots\right).$$

For general α:

$$\frac{(2(1-\alpha)-1)}{\sqrt{(1-\alpha)\alpha}} = \frac{(1-2\alpha)}{\sqrt{\alpha(1-\alpha)}}.$$

Combining terms preserves all α-dependence:

$$T_{\text{stop}} = \frac{2}{\eta}\left(\frac{1}{\sqrt{\beta}} - \frac{1-2\alpha}{\sqrt{\alpha(1-\alpha)}}\right) + \mathcal{O}(\beta^{1/2}).$$

This shows explicit dependence on both α and β. □

D.1.3 Good Arm Dynamics and Variance

Lemma D.3 (Centered Good Arm Dynamics). *Let $\bar{\theta} = \frac{1}{K}\sum_{i=1}^{K}\theta_i$, and define the centered parameters $x_i = \theta_i - \bar{\theta}$ for $i = 1, \ldots, K$. If $p_{K+1} = 1-\alpha$ is small, then one can show*

$$\mathbb{E}[\Delta x_i] \;=\; 0, \quad \mathrm{Var}(\Delta x_i) \;\approx\; \frac{\eta^2}{G}\,\alpha\Big(1 - \frac{\alpha}{K}\Big) \;\times\; \left(\frac{K}{K-1}\right)$$

to account for the negative correlation among the x_i's (due to the constraint $\sum_{i=1}^{K} x_i = 0$). In particular, for large K this factor is close to 1, so effectively

$$\mathrm{Var}(\Delta x_i) \;\approx\; \frac{\eta^2}{G}\,\alpha\Big(1 - \frac{\alpha}{K}\Big).$$

Hence, the vector $\boldsymbol{x} = (x_1, \ldots, x_K)$ experiences an (approximately) isotropic diffusion with effective coefficient

$$D_{\text{eff}} \;\approx\; \frac{\eta^2}{G}\,\alpha\Big(1 - \frac{\alpha}{K}\Big).$$

Proof. When $p_{K+1} \approx 0$, almost all arms in the group are among the K good arms. By symmetry among the K good arms, one deduces that

$$\sum_{i=1}^{K} \Delta x_i \;=\; 0 \quad \text{and} \quad \mathbb{E}[\Delta x_i] = 0.$$

The variance calculation follows from standard batch policy gradient variance formulas. One finds that each $\Delta\theta_i$ is of order $\frac{\eta}{G}$ times a random variable with variance proportional to $\sqrt{\alpha\beta}$. Since $\sum \Delta x_i = 0$, the x_i's have a small negative correlation factor of $\frac{K}{K-1}$ in the variance. For large K, this factor is nearly 1, yielding the claimed approximate variance. □

D.1.4 Upper Bound on Good Arm Collapse

Theorem D.4 (Upper Bound on Collapse Probability)**.** *Define the collapse event*

$$\mathcal{C} = \Big\{ \max_{1 \le i \le K} p_i \;\ge\; \Delta \Big\},$$

for a threshold $\Delta > \frac{1}{K}$. Suppose GRPO runs until $p_{K+1} < \beta$ at time T_{stop}. Under the update variance in Lemma D.3, one obtains

$$\Pr\big(\mathcal{C} \text{ by } T_{\text{stop}}\big) \;\le\; \exp\Big(-C\,\frac{K\,\Delta}{G}\,\ln\big(\tfrac{1-\alpha}{\beta}\big)\Big),$$

where $C > 0$ depends on Δ, K, and α. Hence, as G increases, the probability of collapse decays exponentially in G.

Proof. By Lemma D.1 and Lemma D.2, the bad arm probability reduces below β within $O(\frac{1}{\sqrt{\beta}})$ steps. Condition on $p_{K+1} < \beta$ and apply Lemma D.3 to see that $\boldsymbol{x} = (x_1, \ldots, x_K)$ can be viewed (in a discrete-to-diffusion approximation) as a zero-drift process with variance on the order of $\frac{\eta^2}{G}$. Using standard hitting-time or concentration bounds (e.g. Freedman's inequality or reflection principles for Brownian motion), the probability that $\max_i x_i$ reaches the level $d = \ln(\frac{(K-1)\Delta}{1-\Delta})$ (which corresponds to $\max_i p_i \ge \Delta$) decays exponentially in $\frac{G}{\eta^2}$ times the inverse of the time horizon T_{stop}. Since $T_{\text{stop}} = O(\frac{1}{\sqrt{\beta}})$, the net effect introduces a factor depending on $\ln(\frac{1-\alpha}{\beta})$. Combining constants proves the stated bound. □

D.1.5 Lower Bound on Good Arm Collapse via Hitting Times

Theorem D.5 (Lower Bound on Collapse Probability)**.** *Let $\Delta > \frac{1}{K}$, and define*

$$d \;=\; \ln\Big(\frac{(K-1)\,\Delta}{1-\Delta}\Big).$$

Assume that for t beyond some T_0 (i.e. after the bad arm is mostly suppressed) the good-arm centered parameters $x_i = \theta_i - \bar{\theta}$ satisfy

$$\mathbb{E}[\Delta x_i] = 0, \quad \mathrm{Var}(\Delta x_i) \;\ge\; D_{\text{eff}} \;=\; \frac{\eta^2}{G}\,\alpha\Big(1 - \frac{\alpha}{K}\Big).$$

Then there exists a constant $C_1 > 0$ such that

$$\Pr\Big(\max_{1 \le i \le K} p_i \;\ge\; \Delta \;\text{ by }\; T_{\text{stop}}\Big) \;\ge\; \exp\Big(-C_1\,\frac{G\,d^2}{\eta^2\,\alpha\,(1-\frac{\alpha}{K})\,T_{\text{stop}}}\Big) \sim \exp\left(-C_2\frac{Gd^2}{\eta^2\sqrt{\alpha\beta}}\right).$$

If T_{stop} is on the order of $\frac{1}{\sqrt{\beta}} - \frac{1}{\sqrt{\alpha(1-\alpha)}}$, this becomes

$$\Pr(\mathcal{C}) \;\ge\; \exp\Big(-C_2\,\frac{G\,d^2\,\sqrt{\beta(1-\alpha)}}{\eta^2\,\sqrt{\alpha}\big(1-\frac{\alpha}{K}\big)\big(\sqrt{\alpha(1-\alpha)} - \sqrt{\beta}\big)}\Big).$$

Proof. One-dimensional hitting-time formulas for zero-drift (Brownian-like) processes with variance $D_{\text{eff}}\,t$ imply that the probability a single coordinate $x_i(t)$ reaches $d > 0$ by time T is bounded below by a term $\exp\big(-c\,\frac{d^2}{D_{\text{eff}}\,T}\big)$. Taking a union bound over K good arms affects only polynomial factors (hence not the leading exponent). Since T_{stop} is at most on the order of $\frac{1}{\sqrt{\beta}}$, we substitute T_{stop} and $D_{\text{eff}} = \frac{\eta^2}{G}\,\sqrt{\alpha\beta}$, which yields the exponent $\exp\big(-c\,\frac{G\,d^2\,\sqrt{\beta}}{\eta^2\,\alpha(1-\frac{\alpha}{K})}\big)$. This completes the proof. □

D.1.6 MAIN THEOREM

Theorem D.6 (Combined Bounds on Good Arm Collapse). *Let $\mathcal{C} = \{\max_{1\le i\le K} p_i \ge \Delta\}$ be the collapse event. Suppose that the assumptions of Theorems D.4 and D.5 hold. Then there exist constants C, C_1, $C_2 > 0$ such that:*

$$\exp\Big(-C_1 \frac{G\,(\ln(\frac{(K-1)\Delta}{1-\Delta}))^2}{\eta^2\sqrt{\alpha\beta}}\Big) \;\le\; \Pr(\mathcal{C}) \;\le\; \exp\Big(-C\,\frac{K\,\Delta}{G}\,\ln\Big(\frac{1-\alpha}{\beta}\Big)\Big),$$

where $\Delta > \frac{1}{K}$ is the collapse threshold and $\beta > 0$ is the stopping criterion for p_{K+1}. If $T_{\text{stop}} \sim \frac{1}{\sqrt{\beta}}$, then the lower bound exponent incorporates an additional factor of $\sqrt{\beta}$ in place of T_{stop}.

Proof. Combine Theorem D.4 (upper bound) with Theorem D.5 (lower bound). Both rely on Lemmas D.1, D.2, and D.3 for controlling the time scale T_{stop} and the variance structure of the good arms. The final exponents involve the group size G, the threshold Δ, the initial bad arm probability $1-\alpha$, and the stopping threshold β, producing the stated bounds. □

D.2 ANALYSIS ON RL COLLAPSE TO SFT REGIME

D.2.1 PROBLEM SETUP & NOTATION

- We have a finite set of queries (prompts)
$$\mathcal{S} = \{s_1, s_2, \ldots, s_M\}, \quad M < \infty.$$
- For each s_i, there are N_i *correct* answer paths:
$$\mathcal{A}_i = \{\, a_{i1}, \ldots, a_{iN_i}\}.$$
Let $N = \max_i N_i$ be the maximum number of correct paths for any query.
- A policy $\pi_\theta(a \mid s)$ is parameterized by a (Transformer) neural network with parameters θ. For each s_i, define the *accuracy*:
$$x_i \;=\; \sum_{j=1}^{N_i} \pi_\theta\big(a_{ij} \mid s_i\big).$$
- **Poor Diversity Initialization.** We assume that the initial accuracies $(x_1, \ldots, x_M)$ are drawn i.i.d. from
$$\text{Beta}(\epsilon_x, \epsilon_y)$$
with
$$\epsilon_x, \epsilon_y \;\le\; \tau \ll 1.$$
Thus, with high probability each x_i is extremely close to either 0 or 1, indicating a lack of diversity.

D.2.2 BETA TAIL BOUND

We first restate a refined tail bound for the Beta distribution. This result justifies the claim that most queries start with an accuracy extremely close to 0 or 1.

Lemma D.7 (Beta Tail Bound). *Let $x \sim \text{Beta}(\epsilon_x, \epsilon_y)$ with $\epsilon_x, \epsilon_y \le \tau \ll 1$. Then for every $\varepsilon \in (0, \frac{1}{2})$ there exists a constant $C_1 > 0$ (depending only on τ) such that*

$$\Pr\Big(x \in (\varepsilon,\, 1-\varepsilon)\Big) \;\le\; C_1\,\varepsilon.$$

Equivalently,

$$\Pr\Big(x \le \varepsilon \quad \textit{or} \quad x \ge 1-\varepsilon\Big) \;\ge\; 1 - C_1\,\varepsilon.$$

Proof. Let

$$f(x) = \frac{x^{\epsilon_x - 1}(1-x)^{\epsilon_y - 1}}{B(\epsilon_x, \epsilon_y)}, \qquad 0 < x < 1,$$

be the density of $x \sim \text{Beta}(\epsilon_x, \epsilon_y)$ and

$$B(\epsilon_x, \epsilon_y) = \frac{\Gamma(\epsilon_x)\,\Gamma(\epsilon_y)}{\Gamma(\epsilon_x + \epsilon_y)}.$$

When $\epsilon_x, \epsilon_y \ll 1$, using $\Gamma(z) \sim 1/z$ for $z \ll 1$ we have

$$B(\epsilon_x, \epsilon_y) \sim \frac{1/(\epsilon_x\,\epsilon_y)}{1/(\epsilon_x + \epsilon_y)} \;=\; \frac{\epsilon_x + \epsilon_y}{\epsilon_x\,\epsilon_y}.$$

Thus, the normalization is very large and the density is sharply peaked near 0 and 1.

Denote by

$$I(x) = \frac{1}{B(\epsilon_x, \epsilon_y)} \int_0^x t^{\epsilon_x - 1}(1-t)^{\epsilon_y - 1}\,dt,$$

the cumulative distribution function. Standard estimates for the Beta function (Abramowitz & Stegun, 1964) show that for $\varepsilon \in (0, 1/2)$ one may choose a constant $C_1 > 0$ (depending only on τ) so that

$$F(\varepsilon) \;=\; I(\varepsilon) \;\geq\; 1 - C_1\,\varepsilon,$$

and by symmetry

$$1 - F(1-\varepsilon) \;\geq\; 1 - C_1\,\varepsilon.$$

It follows that the total tail probability satisfies

$$\Pr\Big(x \leq \varepsilon \text{ or } x \geq 1-\varepsilon\Big) \;=\; F(\varepsilon) + \Big[1 - F(1-\varepsilon)\Big] \;\geq\; 1 - C_1\,\varepsilon,$$

or equivalently,

$$\Pr\Big(x \in (\varepsilon, 1-\varepsilon)\Big) \;\leq\; C_1\,\varepsilon.$$

This completes the proof. □

D.2.3 Main Theorem

We now state our main result. It shows that under GRPO training the policy for each query eventually collapses to a single dominant correct answer path—thus effectively entering a supervised fine-tuning (SFT) regime. Importantly, the only dependence on the initialization appears via the parameter τ.

Theorem D.8 (RL Collapse to SFT Regime)**.** *Assume that the GRPO training is applied independently to each query $s_i \in \mathcal{S}$ (with $|\mathcal{S}| = M < \infty$), where for each query the set of correct answer paths is identified with the K good arms of a $(K+1)$-armed bandit. Suppose that the initial accuracy*

$$x_i \;=\; \sum_{j=1}^{N_i} \pi_\theta(a_{ij} \mid s_i)$$

is drawn i.i.d. from $\text{Beta}(\epsilon_x, \epsilon_y)$ *with*

$$\epsilon_x, \epsilon_y \;\leq\; \tau \ll 1.$$

Then by Lemma D.7 most queries are initialized with either $x_i \leq c\tau$ or $x_i \geq 1 - c\tau$ for some constant $c > 0$. In the low-accuracy case $x_i \leq c\tau$, the total probability α on the K good arms satisfies $\alpha \leq c\tau$. By applying Theorem D.6 (Combined Bounds on Good Arm Collapse) with the substitution $\alpha \leq c\tau$ and absorbing the stopping criterion β into the universal constants, we deduce that the collapse event

$$\mathcal{C}_i \;=\; \Big\{ \max_{1 \leq j \leq K} \pi_\theta(a_{ij} \mid s_i) \geq \Delta \Big\}, \quad \Delta > \tfrac{1}{K},$$

satisfies

$$\Pr(\mathcal{C}_i) \;\geq\; \exp\!\left(-\,\widetilde{C}_1\, \frac{G\left(\ln\!\left(\frac{(K-1)\Delta}{1-\Delta}\right)\right)^2}{\eta^2\,\tau} \right),$$

where $\widetilde{C}_1 > 0$ is a constant independent of τ. (In the high-accuracy case $x_i \geq 1 - c\tau$ the policy is already collapsed.) Hence, by a union bound over all M queries, the probability that the policy collapses to a single dominant correct answer path for every query (i.e., that the RL training degenerates into an SFT regime) satisfies

$$\Pr\Big(\bigcap_{i=1}^{M} \mathcal{C}_i\Big) \;\geq\; 1 \;-\; M\, \left(1 - \exp\!\left(-\,\widetilde{C}_1\, \frac{G\left(\ln\!\left(\frac{(K-1)\Delta}{1-\Delta}\right)\right)^2}{\eta^2\,\tau} \right) \right).$$

gIn particular, for sufficiently small τ the collapse occurs with high probability.

D.3 COLLAPSE ANALYSIS OF REINFORCE

In this section, we provide an analysis of the diversity collapse for REINFORCE update in the multi-armed bandit settings (both discrete and continuous). Our main results are two theorems, one for the discrete bandit (**Theorem D.10**) and one for the continuous bandit (**Theorem D.11**).

D.3.1 DISCRETE BANDIT SETTING

Setup. We consider a $(K+1)$-armed bandit, with arms $\{1, 2, \ldots, K, K+1\}$. Arms $1, \ldots, K$ are "good," each yielding reward 1, and arm $K+1$ is "bad," yielding reward 0. We use a softmax parameterization:

$$p_i \;=\; \frac{e^{\theta_i}}{\sum_{j=1}^{K+1} e^{\theta_j}}, \quad i = 1, \ldots, K+1.$$

At each step:

1. We sample an arm I_t according to $p(\cdot) = (p_1, \ldots, p_{K+1})$.
2. The reward is $r_t = 1$ if $I_t \leq K$ (a good arm), and $r_t = 0$ otherwise.
3. We update using policy gradient from REINFORCE

$$\theta_i \;\leftarrow\; \theta_i \;+\; \eta\, r_t \left(\mathbf{1}\{I_t = i\} \;-\; p_i\right), \quad i = 1, \ldots, K+1,$$

   where $\eta > 0$ is the step size.
4. We stop once the bad arm's probability mass p_{K+1} drops below some fixed $\beta > 0$. Denote this (random) stopping time by

$$T_{\text{stop}} \;=\; \min\Big\{t \;\mid\; p_{K+1}^{(t)} \;<\; \beta\Big\}.$$

Let $\alpha \in (0, 1)$ be the initial total probability mass on the K good arms so that $p_i(0) = \frac{\alpha}{K}$ for $i = 1, \ldots, K$ and $p_{K+1}(0) = 1 - \alpha$. Our goal is to show that, once the bad arm is essentially removed (i.e. $p_{K+1} < \beta$), the policy can *collapse* among the remaining good arms with significant probability; in other words, it eventually concentrates almost all of its mass on one of the K arms.

Formally, define

$$p_i^{(g)} \;=\; \frac{e^{\theta_i}}{\sum_{j=1}^{K} e^{\theta_j}} \quad \text{for } i = 1, \ldots, K.$$

This $p^{(g)}(\cdot)$ is simply the softmax restricted to the good arms. We say that a *collapse* event occurs if

$$\max_{1 \leq i \leq K} p_i^{(g)} \;\geq\; \Delta,$$

for some fixed threshold $\Delta \in (0, 1)$. We will show that with nontrivial probability, the policy has already collapsed among the good arms *by the time T_{stop} at which the bad arm's probability mass goes below β*.

Dynamics of the Bad Arm. We first estimate the time T_{stop} needed for p_{K+1} to drop below β. The update for θ_{K+1} at step t is:

$$\Delta\theta_{K+1} \;=\; \eta\, r_t \big(\mathbf{1}\{I_t = K+1\} - p_{K+1}\big).$$

Since the reward is $r_t = 0$ whenever $I_t = K+1$, the actual increment can be broken down into two cases each step:

$$\Delta\theta_{K+1} \;=\; \begin{cases} 0, & \text{if } I_t = K+1, \\ -\eta\, p_{K+1}, & \text{if } I_t \neq K+1. \end{cases}$$

Hence the expected increment in θ_{K+1} at step t is

$$\mathbb{E}[\Delta\theta_{K+1} \mid \text{state at } t] \;=\; -\,\eta\, p_{K+1} \sum_{i=1}^{K} p_i \;=\; -\,\eta\, p_{K+1}\,\big(1 - p_{K+1}\big).$$

This negative drift in θ_{K+1} causes p_{K+1} to decay. A standard approximation (treating the update as continuous) suggests

$$p_{K+1}(t+1) \;\approx\; p_{K+1}(t)\,\exp\big(-\eta\,(1-p_{K+1}(t))\big),$$

which is close to a nearly-exponential decay when p_{K+1} is not too large.

To make this rigorous, one can apply a concentration argument (e.g. Freedman's inequality or Azuma–Hoeffding) to the sum of increments

$$\Delta\theta_{K+1}(1) + \Delta\theta_{K+1}(2) + \cdots + \Delta\theta_{K+1}(t).$$

This ensures that with high probability, p_{K+1} indeed decreases from $(1-\alpha)$ to β within a time on the order of

$$\frac{1}{\eta}\,\ln\Big(\frac{1-\alpha}{\beta}\Big).$$

Thus we get the following high-probability bound:

$$T_{\text{stop}} \;=\; \min\{t : p_{K+1}(t) < \beta\} \;\lesssim\; \frac{1}{\eta}\,\ln\Big(\frac{1-\alpha}{\beta}\Big). \tag{9}$$

Centering the Good-Arms Parameters. Once p_{K+1} is small, most draws are from the K good arms. It is helpful to define

$$\bar\theta \;=\; \frac{1}{K}\sum_{i=1}^{K}\theta_i, \quad x_i \;=\; \theta_i - \bar\theta, \quad i = 1,\dots,K.$$

Clearly, $\sum_{i=1}^{K} x_i = 0$, so $x = (x_1,\dots,x_K)$ lives in a $(K-1)$-dimensional subspace $\{x : \sum_i x_i = 0\}$. We also define

$$p_i^{(g)} \;=\; \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}},$$

which is just the good-arm softmax re-expressed in terms of x. A collapse event means $\max_i p_i^{(g)} > \Delta$, i.e., equivalently $x_i > \ln(\Delta/(1-\Delta))$ for some i. Let us set

$$\Delta_\theta \;=\; \ln\Big(\tfrac{\Delta}{1-\Delta}\Big),$$

and denote by $\mathcal{D}$ the set

$$\mathcal{D} \;=\; \Big\{x \in \mathbb{R}^K : \max_{1\le i\le K} x_i < \Delta_\theta,\; \sum_{i=1}^{K} x_i = 0\Big\}.$$

We interpret $\mathcal{D}$ as the "no-collapse" domain in the $(x_1,\dots,x_K)$-space. Once $x \notin \mathcal{D}$, we have a collapse.

Martingale CLT and Diffusion Approximation. Next, we show that the evolution of $x_i = \theta_i - \bar\theta$ behaves roughly like a $(K-1)$-dimensional Brownian motion on moderate time intervals, provided η is chosen so that $\eta^2 N$ stays of order 1 over relevant horizons N.

Observe that the update to each θ_i has magnitude of order η, since $r_t \in \{0,1\}$ and $|\,\mathbf{1}\{I_t = i\} - p_i| \le 1$. Hence

$$\mathrm{Var}\big(\Delta\theta_i\big) \;=\; \mathcal{O}(\eta^2).$$

Moreover,

$$\Delta x_i \;=\; \Delta\theta_i \;-\; \frac{1}{K}\sum_{j=1}^{K}\Delta\theta_j.$$

When the probability mass is mostly on the K good arms and not overly concentrated on any single good arm, p_i is roughly $1/K$ for each $i = 1,\dots,K$. In that regime, $(\Delta x_1,\dots,\Delta x_K)$ acts like a mean-0 increment with variance on the order of η^2. By a standard martingale central limit argument (Karatzas & Shreve, 1991), summing Δx_i over N steps is close in distribution to a $(K-1)$-dimensional Gaussian with covariance $\propto N\,\eta^2$.

Scaling time continuously, one views $t = N\eta^{-2}$ (or ensures $\eta^2 N$ is $O(1)$), then $x(\cdot)$ converges to a diffusion with generator $\frac{1}{2}D_{\text{eff}}\Delta$, for some *effective diffusion coefficient* $D_{\text{eff}} > 0$. One can regard

$$D_{\text{eff}} \;=\; \alpha\,C\,\eta^2$$

for some constant $C > 0$ (depending on K), and one should keep track of the explicit factor α (the initial probability mass on good arms) to avoid absorbing it as a mere constant.

Exit-Time Bounds via Elliptic PDEs. In order to show that $x(t)$ (the centered good-arm parameters) escapes the domain $\mathcal{D}$ by a certain time, we rely on a standard exit-time bound for Brownian-like diffusions in a bounded domain. Concretely, let X_t be a $(K-1)$-dimensional diffusion with generator

$$\mathcal{L} \;=\; \tfrac{1}{2}\, D_{\text{eff}}\, \Delta \quad \text{on a bounded domain } \mathcal{D}.$$

Define $\tau_{\mathcal{D}} = \inf\{t \geq 0 : X_t \notin \mathcal{D}\}$. Then one has an exponentially small upper bound on the probability that $\tau_{\mathcal{D}}$ is large. The heart of the argument comes from solving or bounding the heat equation with Dirichlet boundary. For completeness:

Lemma D.9 (Exit-Time Bound)**.** *Let X_t be the $(K-1)$-dimensional diffusion solving $dX_t = \sqrt{D_{\text{eff}}}\, dW_t$, starting in a bounded convex domain $\mathcal{D} \subset \mathbb{R}^{K-1}$ with smooth boundary. Define $\tau_{\mathcal{D}} = \inf\{t \geq 0 : X_t \notin \mathcal{D}\}$. Then there exist constants $c_1 > 0$ and $R > 0$ (depending on $\mathcal{D}$ and K) such that for all $T > 0$,*

$$\mathbb{P}\big(\tau_{\mathcal{D}} > T\big) \;\leq\; \exp\Big(-\,c_1\, \frac{D_{\text{eff}}}{R^2}\, T\Big).$$

Proof. **Step 1.** Consider the backward heat equation with Dirichlet boundary condition on $\mathcal{D}$:

$$\begin{cases} \partial_t u \;=\; \tfrac{1}{2}\, D_{\text{eff}}\, \Delta u, & x \in \mathcal{D}, \\ u(x,t) \;=\; 0, & x \in \partial\mathcal{D}, \\ u(x,0) \;=\; 1, & x \in \mathcal{D}. \end{cases}$$

By expanding u in an orthonormal basis of Dirichlet eigenfunctions $\{\phi_k\}$ of $-\Delta$ with eigenvalues $\{\lambda_k\} > 0$, we have

$$-\Delta\phi_k \;=\; \lambda_k\, \phi_k, \quad \phi_k\Big|_{\partial\mathcal{D}} = 0.$$

Thus

$$u(x,t) \;=\; \sum_{k=1}^{\infty} \Big(\int_{\mathcal{D}} \phi_k(y)\, dy\Big)\, \phi_k(x)\, \exp\big(-\tfrac{1}{2}\, D_{\text{eff}}\, \lambda_k\, t\big).$$

Since $\lambda_1 > 0$ is the smallest (principal) eigenvalue, and ϕ_1 can be chosen nonnegative on $\mathcal{D}$, there exists a uniform positive lower bound on $\phi_1(x)$ for $x \in \mathcal{D}$. Consequently,

$$u(x,t) \;\leq\; \big[\sup_{y\in\mathcal{D}} \phi_1(y)\big]\; \phi_1(x)\; \exp\big(-\tfrac{1}{2}\, D_{\text{eff}} \lambda_1\, t\big).$$

After a suitable normalization of ϕ_1, we conclude

$$u(x,t) \;\leq\; \exp\big(-\tfrac{1}{2}\, D_{\text{eff}} \lambda_1\, t\big).$$

Step 2. By optional stopping (applied to the martingale $u(X_{t\wedge\tau_{\mathcal{D}}},\, t \wedge \tau_{\mathcal{D}})$), we get

$$u(x,0) \;=\; 1 \;=\; \mathbb{E}_x\big[u\big(X_{t\wedge\tau_{\mathcal{D}}}, t\wedge\tau_{\mathcal{D}}\big)\big] \;\geq\; \mathbb{P}_x(\tau_{\mathcal{D}} > t)\; \min_{y\in\mathcal{D}} u(y,t).$$

Since $u(\cdot,t)$ is nonnegative, we obtain

$$\mathbb{P}_x(\tau_{\mathcal{D}} > t) \;\leq\; \frac{\max_{y\in\mathcal{D}} u(y,0)}{\min_{y\in\mathcal{D}} u(y,t)} \;\leq\; \exp\big(-\tfrac{1}{2}\, D_{\text{eff}} \lambda_1\, t\big).$$

Step 3. By a domain-geometric (Faber–Krahn type) bound or explicit PDE estimates for bounded convex sets in $\mathbb{R}^{K-1}$, one obtains $\lambda_1 \geq c_1'/R^2$ for some $c_1' > 0$. Let R be, for instance, the minimal radius of an inscribed ball in $\mathcal{D}$. Combining, we get

$$\mathbb{P}_x(\tau_{\mathcal{D}} > t) \;\leq\; \exp\Big(-\,\tfrac{1}{2}\, c_1'\, \tfrac{D_{\text{eff}}}{R^2}\; t\Big).$$

Renaming $c_1 = \frac{1}{2}\, c_1'$, the result follows. □

In our application, $\mathcal{D} = \{\, x : \max_i x_i < \Delta_\theta, \sum_i x_i = 0\}$ is an intersection of K half-spaces, hence a bounded convex domain within that $(K-1)$-dimensional subspace. We let $R > 0$ be a suitable diameter or inradius for $\mathcal{D}$.

Main Result for the Discrete Bandit.

Theorem D.10 (Discrete Bandit Collapse)**.** *Consider the $(K+1)$-armed bandit with K good arms and 1 bad arm, using the softmax policy gradient update. Let $\alpha \in (0,1)$ be the **initial** total probability mass on the K good arms. Fix parameters $\beta > 0$ and $\Delta \in (0,1)$. Suppose the step size η is chosen such that $\eta^2 N$ remains of order 1 over the relevant horizons N. Then there exists a constant $c > 0$ (depending on K and the geometry of $\mathcal{D}$) such that:*

1. *With high probability, the probability mass on the bad arm decays to below β by time*

$$T_{\text{stop}} \lesssim \frac{1}{\eta} \ln\Big(\frac{1-\alpha}{\beta}\Big).$$

2. *With the diffusion approximation in the good-arms subspace, the probability of* not *collapsing (i.e. staying inside $\mathcal{D}$) for all times up to T_{stop} is at most*

$$\exp\Big(-c\,\frac{D_{\text{eff}}}{(\Delta_\theta)^2}\,T_{\text{stop}}\Big) \approx \exp\Big(-c\,\alpha\,\eta\,\ln\big(\tfrac{1-\alpha}{\beta}\big)\Big),$$

where $D_{\text{eff}} \approx \alpha\,\eta^2$ and $\Delta_\theta = \ln(\Delta/(1-\Delta))$. Consequently,

$$\mathbb{P}\big(\tau_{\mathcal{D}} \le T_{\text{stop}}\big) \ge 1 - \exp\Big(-c\,\alpha\,\eta\,\ln\big(\tfrac{1-\alpha}{\beta}\big)\Big),$$

implying that with non-negligible probability, the policy has collapsed *onto a single good arm (i.e. $\max_i p_i^{(g)} > \Delta$)* by *the time the bad arm is essentially removed.*

Proof. **(1) Time to remove the bad arm.** From the update rule for θ_{K+1}, we see that every time a good arm is selected (which happens with probability $1 - p_{K+1}$), θ_{K+1} decreases by approximately $\eta\,p_{K+1}$. Summing these increments shows that p_{K+1} decays nearly exponentially from $(1-\alpha)$ to below β. A standard Azuma–Hoeffding or Freedman-style inequality applied to the martingale increments ensures that this decay happens with high probability in time on the order of $\frac{1}{\eta}\ln\big(\frac{1-\alpha}{\beta}\big)$; see equation 9.

(2) Collapse among the good arms. Once p_{K+1} is sufficiently small, most updates occur on the K good arms. In that regime, one can approximate $p_i^{(g)} \approx 1/K$ for each good arm, implying the increments Δx_i have mean near zero and variance of order η^2. By the martingale CLT, $x(t)$ behaves approximately like a Brownian motion with diffusion $D_{\text{eff}} \approx \alpha\,\eta^2$. Applying Lemma D.9 on the "no-collapse" domain $\mathcal{D}$, we see that the probability of *remaining* inside $\mathcal{D}$ up to T_{stop} is at most

$$\exp\Big(-c''\,\frac{D_{\text{eff}}}{(\Delta_\theta)^2}\,T_{\text{stop}}\Big) \approx \exp\Big(-c''\,\frac{\alpha\,\eta^2}{(\Delta_\theta)^2}\,\frac{1}{\eta}\,\ln\big(\tfrac{1-\alpha}{\beta}\big)\Big) = \exp\Big(-c''\,\alpha\,\eta\,\ln\big(\tfrac{1-\alpha}{\beta}\big)\Big).$$

Hence, with probability at least $1 - \exp(\cdots)$, the process exits $\mathcal{D}$ (i.e. experiences $\max_i p_i^{(g)} > \Delta$) before T_{stop}. This exit event is precisely the definition of *collapse*. □

D.3.2 Continuous Bandit Setting

Setup. In many cases, the number of correct traces is exponential to the input size, and in those cases, discrete bandit setting may not be proper to describe the situations. We now consider a continuum-armed bandit with action space $[0,1]$. Let $G \subseteq [0,1]$ be a measurable "good" set (reward 1) and its complement $[0,1] \setminus G$ be "bad" (reward 0). We use a softmax policy parameterized by $v \in \mathbb{R}^d$ and a feature map $\Phi : [0,1] \to \mathbb{R}^d$. Concretely,

$$p_v(a) = \frac{\exp\big(v^\top \Phi(a)\big)}{\displaystyle\int_0^1 \exp\big(v^\top \Phi(x)\big)\,dx},$$

and at each iteration:

1. Sample $a \sim p_v(\cdot)$,
2. Reward $r(a) = 1$ if $a \in G$, else 0,

3. Update
$$v \leftarrow v + \eta\, r(a)\,\big(\Phi(a) - \mathbb{E}_{p_v}[\Phi(\cdot)]\big).$$
4. Stop once the bad-region mass $\int_{[0,1]\setminus G} p_v(a)\,da$ is below β. Denote this stopping time by T_{stop}.

Let $\alpha = \int_G p_v(a)\,da$ be the initial mass on the good set G. We wish to show that once $[0,1] \setminus G$ is nearly pruned ($< \beta$ mass), the distribution in G can collapse onto a small subregion of G with nontrivial probability.

Dynamics of the Bad Region. Define
$$m_B(v) \;=\; \int_{[0,1]\setminus G} p_v(a)\,da.$$
Initially $m_B(v(0)) = 1 - \alpha$. The gradient update for v on $[0,1] \setminus G$ receives reward 0. Hence the net effect is to push v so as to *decrease* $m_B(v)$ in expectation. A concentration argument again shows an approximate exponential decay:
$$m_B\big(v(t)\big) \;\lesssim\; (1-\alpha)\,\exp\!\big(-\eta\, t\big),$$
so the time to reach $m_B(v) \le \beta$ is on the order of
$$T_{\text{stop}} \;\lesssim\; \frac{1}{\eta}\,\ln\!\Big(\frac{1-\alpha}{\beta}\Big).$$

Diffusion Approximation Within G. Decompose v as (v_G, v_B) according to features supported on G versus $[0,1] \setminus G$. Once $m_B < \beta$, the sampling is mostly from G, making the updates to v_G predominantly come from that region. If $p_v(\cdot)$ is near-uniform on G, then $\mathbb{E}[\Phi(a) \mid a \in G] \approx \bar{\Phi}_G$ and the *mean* increment of v_G is small, while the *variance* remains of order η^2. As in the discrete case, a martingale CLT argument shows
$$v_G(t) \;\approx\; \text{Brownian motion with diffusion } D_{\text{eff}} \;\approx\; \alpha\,\eta^2\,\Sigma_G,$$
where Σ_G depends on the feature covariance over G. Again, the factor α is kept explicit.

Defining Collapse in the Continuous Setting. Inside G, define the *restricted policy*
$$\tilde{p}_v(a) \;=\; \frac{p_v(a)}{\int_G p_v(x)\,dx}, \quad a \in G.$$
We say *collapse* occurs if
$$\sup_{a\in G} \tilde{p}_v(a) > \Delta,$$
for some $\Delta > 1/\mu(G)$ (assuming G has positive Lebesgue measure $\mu(G)$). Intuitively, if $\tilde{p}_v(\cdot)$ becomes very peaked on a small subset of G, the exploration collapses.

Let $U \subset \mathbb{R}^d$ be the set of v_G for which $\sup_{a\in G} \tilde{p}_v(a) \le \Delta$. Define $\tau_U = \inf\{t : v_G(t) \notin U\}$, the first exit time from the "no-collapse" set U. As before, the PDE-based exit-time bound implies that for a diffusion $dX_t = \sqrt{D_{\text{eff}}}\,dW_t$,
$$\mathbb{P}\big(\tau_U > T\big) \;\le\; \exp\!\Big(-c_1\,\tfrac{D_{\text{eff}}}{R^2}\,T\Big),$$
provided U is bounded and sufficiently regular. Here, R is a diameter or inradius of U.

Main Result for the Continuous Bandit.

Theorem D.11 (Continuous Bandit Collapse)**.** *Consider the continuum-armed bandit on $[0,1]$ with reward 1 on a measurable subset $G \subset [0,1]$ and reward 0 elsewhere. Let $\alpha = \int_G p_v(a)\,da$ be the initial mass on G. Suppose $\beta > 0$ and choose a collapse threshold $\Delta > 1/\mu(G)$. Assume the step size η is chosen so that $\eta^2 N$ remains of order 1 over relevant horizons N. Then there is a constant $c > 0$ (depending on U and the feature covariance on G) such that:*

1. *With high probability, the mass on* $[0,1] \setminus G$ *decays below* β *by time*

$$T_{\text{stop}} \lesssim \frac{1}{\eta} \ln\Big(\frac{1-\alpha}{\beta}\Big).$$

2. *The probability of remaining in the no-collapse set* U *up to* T_{stop} *is exponentially small in* T_{stop}*:*

$$\mathbb{P}\big(\tau_U > T_{\text{stop}}\big) \le \exp\Big(-c\,\frac{D_{\text{eff}}\,T_{\text{stop}}}{(\Delta'_\theta)^2}\Big),$$

where $D_{\text{eff}} \approx \alpha\,\eta^2\,\lambda_{\max}(\Sigma_G)$*, and* Δ'_θ *depends on* Δ *and* Φ*. Therefore,*

$$\mathbb{P}\big(\tau_U \le T_{\text{stop}}\big) \ge 1 - \exp\Big(-c\,\alpha\,\eta\,\ln\big(\tfrac{1-\alpha}{\beta}\big)\Big),$$

implying that with non-negligible probability, the policy collapses *onto a narrow subregion of* G *by the time the bad region is essentially pruned.*

Proof. **(1) Pruning the bad region.** Because actions in $[0,1]\setminus G$ yield zero reward, the corresponding coordinates of v have negative drift in expectation, thus $m_B(v(t))$ decays nearly exponentially from $(1-\alpha)$ to β. A high-probability concentration argument shows

$$T_{\text{stop}} = \min\{\,t : m_B(v(t)) < \beta\} \lesssim \frac{1}{\eta} \ln\Big(\frac{1-\alpha}{\beta}\Big).$$

(2) Collapse within G**.** Once $m_B < \beta$, sampling is mainly from G, so v_G accumulates small-mean, $\mathcal{O}(\eta^2)$-variance increments. By the martingale CLT, $v_G(t)$ is well-approximated by a diffusion $dv_G = \sqrt{D_{\text{eff}}}\,dW_t$. Let U be the set of v_G that keep $\max_{a\in G}\tilde{p}_v(a) \le \Delta$. Applying the exit-time argument (as in Lemma D.9 but now in $\mathbb{R}^d$) shows

$$\mathbb{P}\big(\tau_U > T_{\text{stop}}\big) \le \exp\Big(-c_1\,\tfrac{D_{\text{eff}}}{(\Delta'_\theta)^2}\,T_{\text{stop}}\Big) \approx \exp\Big(-c_1\,\alpha\,\eta\,\ln\big(\tfrac{1-\alpha}{\beta}\big)\Big).$$

Hence, with probability at least $1-\exp(\cdots)$, we exit U (i.e. have $\sup_{a\in G}\tilde{p}_v(a) > \Delta$) by T_{stop}, which is the desired collapse event. □

Thus, in both discrete and continuous settings, simple policy-gradient updates exhibit a pronounced "diversity collapse" behavior: once the bad actions are pruned away, the algorithm can rapidly concentrate all remaining probability mass onto a single action (or a very narrow region), despite the existence of multiple equally good actions. Crucially, the probability of collapse scales exponentially in $\alpha\,\eta\,\ln\big(\frac{1-\alpha}{\beta}\big)$, so the effect is *not* negligible when α, β are held fixed and η is suitably chosen.

E MEASURING DIVERSITY OF TRACES

We measure the *diversity* of the 100 sampled rollouts of Gemma-2-2B across GSM8k test. We measure diversity in terms of 3 different measures.

Output Diversity The cardinality or number of unique answers in the set of all model outputs $|\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}|$ over the total number of traces.

Operation Diversity In GSM8k, each intermediate step consists of basic arithmetic operations, e.g. 5 + 3 = 8. We may simply map each of the rollouts to the sequence of arithmetic operations the model steps through, i.e. $r_i \to [o_1, o_2, \ldots, o_t]$. This mapping is extracted by code. Then, given this set, we measure unique sequence of operations over the number of total rollouts.

Semantic Diversity We measure the similarity of trace using cosine similarities between the text-embeddings (Bilmes, 2022; Yu et al., 2023).

In Figure 7, we plot the histogram of these metrics evaluated over the GSM8k test set. Interestingly, we observe a clear trend where the answer diversity decreases and the semantic diveristy and the diversity of approaches in terms of sequence of operations collapses.

E.1 DOES TEMPERATURE INCREASE DIVERSITY?

Temperature does increase diversity, however also increases the chances of sampling outlier answers as we saw in Section 5.

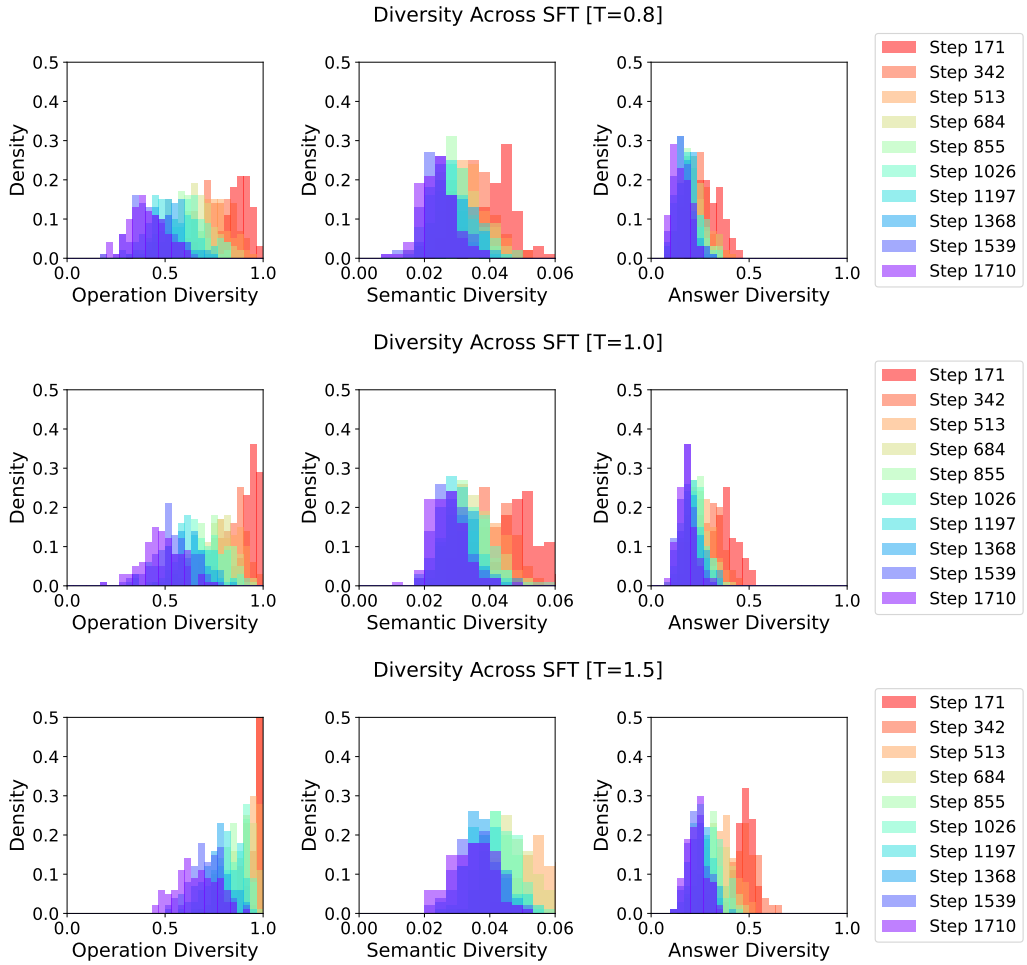

Figure 7: Diversity of Rollouts Sampled w/ Temperature=[0.8, 1.0, 1.5] of Gemma-2-2B SFT checkpoints on GSM8k

E.2 How well do token-level diverse decoding strategies compare with optimal strategy with oracle?

Decoding Strategy	Pass@2	Pass@4	Pass@8
Naive	0.565	0.666	0.760
Nucleus	0.566	0.668	0.757
Min-p	0.566	0.668	0.760
Top-k	0.563	0.666	0.756
Top-k w/Oracle	0.760	0.832	0.901

Table 1: Best Pass@k of Sampling Strategies for Qwen-2.5-0.5B over SFT checkpoints

Decoding Strategy	Pass@2	Pass@4	Pass@8
Naive	0.547	0.648	0.737
Nucleus	0.528	0.617	0.694
Min-p	0.550	0.655	0.744
Top-k	0.538	0.646	0.738
Top-k w/Oracle	0.730	0.814	0.878

Table 2: Pass@k of Sampling Strategies for Qwen-2.5-0.5B at Last SFT Checkpoint

Hyperparameter Tuning Details We grid search for optimal temperature for all baselines over $T = [0.8, 1.0, 1.2, 1.5, 1.8]$. For nucleus, we choose the best cutoff threshold between $[0.8, 0.9, 0.95]$. For min-p, we choose the best probability threshold between $[0.01, 0.05, 0.1]$. For tokenwise top-k, we choose best k between $[12, 25, 50]$.

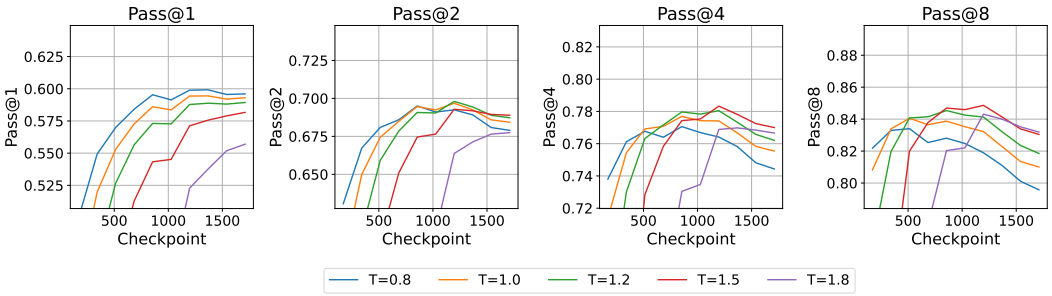

Figure 8: Pass@k of Gemma-2-2B GSM8k Naive Sampling with Replacement

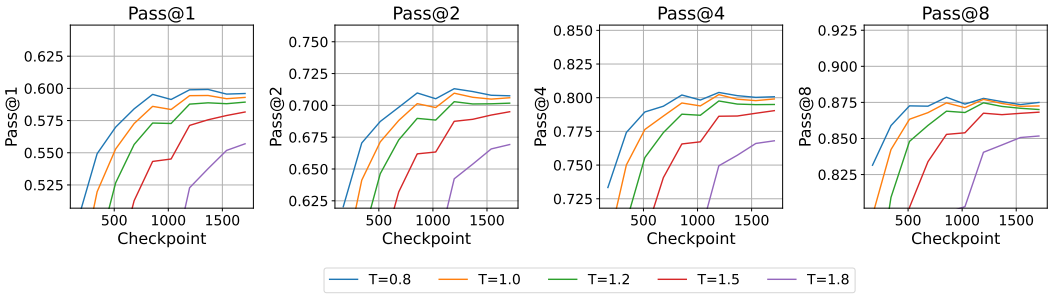

Figure 9: Pass@k of Gemma-2-2B GSM8k Oracle Sampling without Replacement

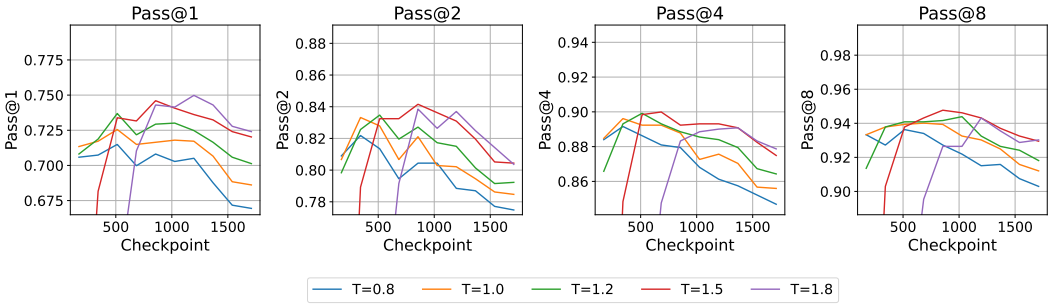

Figure 10: Pass@k of Gemma-2-2B GSM8k Oracle Top K Sampling

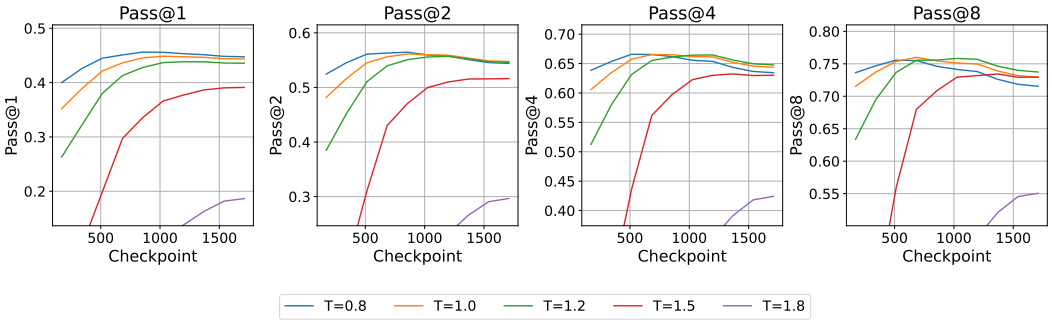

Figure 11: Pass@k of Qwen-2.5-0.5B GSM8k Naive Sampling with Replacement

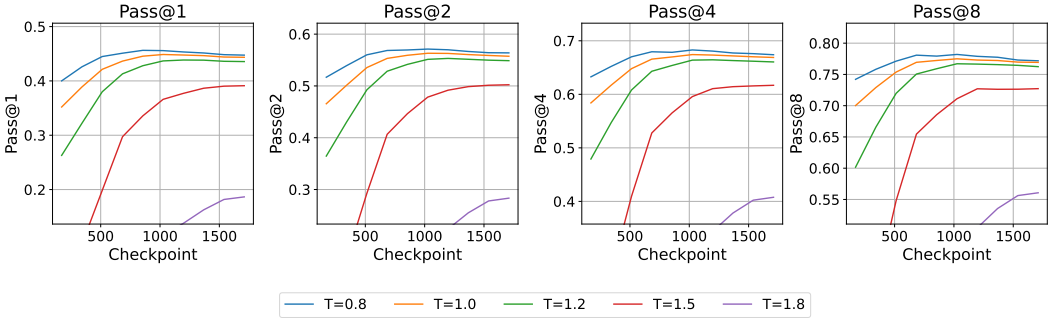

Figure 12: Pass@k of Qwen-2.5-0.5B GSM8k Oracle Sampling without Replacement

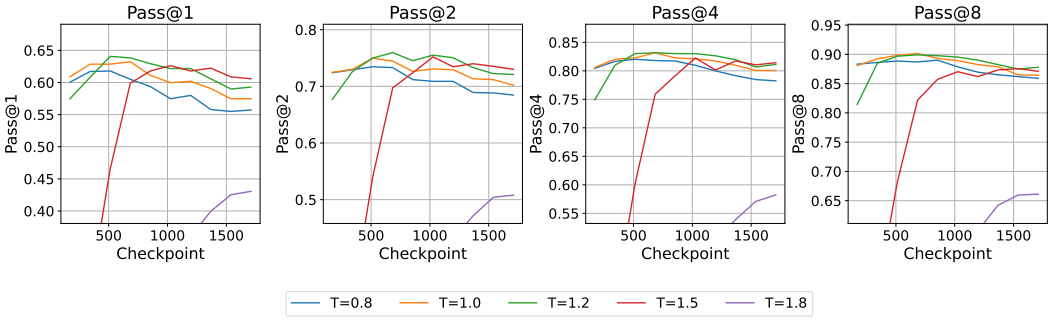

Figure 13: Pass@k of Qwen-2.5-0.5B GSM8k Oracle Top K Sampling

E.3 WHY MIGHT INCREASING TEMPERATURE MAY NOT LEAD TO A "MEANINGFUL" INCREASE IN DIVERSITY?

While the *variance* of the distribution of expected accuracies decreases after temperature scaling (Good for Pass@k), the *overall expected accuracy* also monotonically decreases.

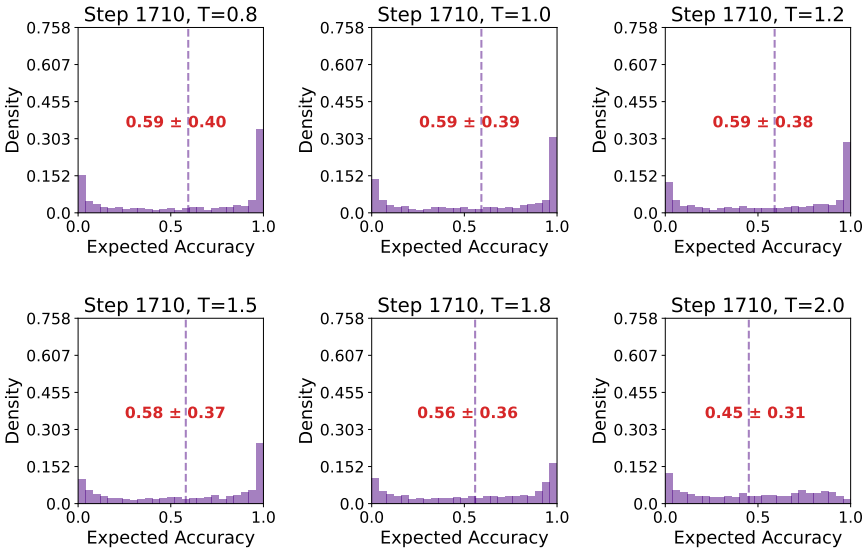

Figure 14: Histogram of ρ_x over GSM8k of Gemma-2-2B Sampled with Temperatures 0.8 to 2.0 after SFT Step 1710

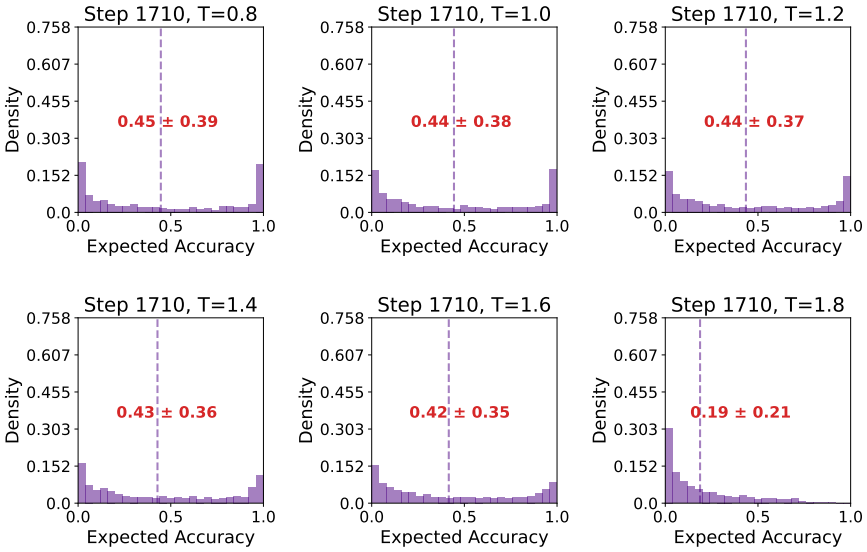

Figure 15: Histogram of ρ_x over GSM8k of Qwen-2.5-0.5B Sampled with Temperatures 0.8 to 1.8 after SFT Step 1710

E.4 WEIGHT INTERPOLATION

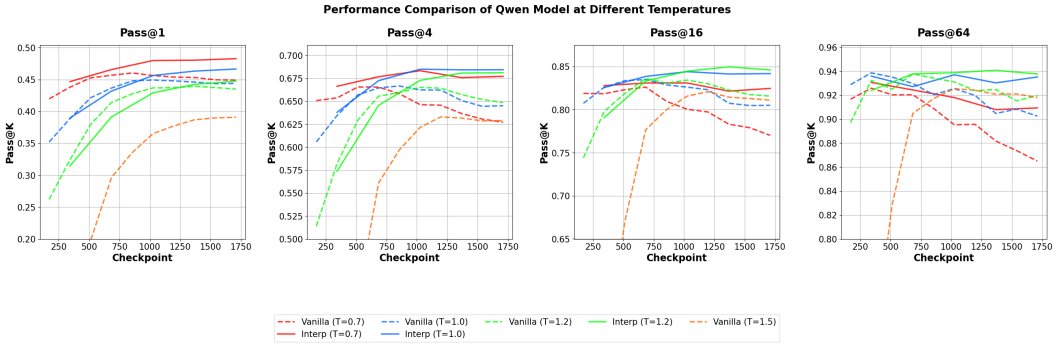

Figure 16: Pass @k for Qwen-2.5-0.5B GSM8k weight interpolation