SED-SFT: Selectively Encouraging Diversity in Supervised Fine-Tuning

Yijie Chen*,**Yijin Liu*** **and Fandong Meng**
WeChat AI, Tencent Inc, China
{kantichen, yijinliu, fandongmeng}@tencent.com

Abstract

Supervised Fine-Tuning (SFT) followed by Reinforcement Learning (RL) has emerged as the standard post-training paradigm for large language models (LLMs). However, the conventional SFT process, driven by Cross-Entropy (CE) loss, often induces mode collapse, where models over-concentrate on specific response patterns. This lack of distributional diversity severely restricts the exploration efficiency required for subsequent RL. While recent studies have attempted to improve SFT by replacing the CE loss, aiming to preserve diversity or refine the update policy, they fail to adequately balance diversity and accuracy, thereby yielding suboptimal performance after RL. To address the mode collapse problem, we propose SED-SFT, which adaptively encourages diversity based on the token exploration space. This framework introduces a selective entropy regularization term with a selective masking mechanism into the optimization objective. Extensive experiments across eight mathematical benchmarks demonstrate that SED-SFT significantly enhances generation diversity with a negligible computational overhead increase compared with CE loss, yielding average improvements of 2.06 and 1.20 points in subsequent RL performance over standard CE-based baselines on Llama-3.2-3B-Instruct and Qwen2.5-Math-7B-Instruct, respectively. [1]

1 Introduction

The prevailing post-training paradigm for Large Language Models (LLMs) typically involves Supervised Fine-Tuning (SFT) followed by Reinforcement Learning (RL). SFT primarily serves to align models with human instructions and enhance specialized capabilities cost-effectively. The mainstream research in SFT has focused on data engineering (Gunasekar et al., 2023; Lambert et al., 2024; Zhou et al., 2023). However, the standard Cross-Entropy (CE) loss mechanism, which pushes probability mass towards the target label, implicitly suppresses diversity. This reduction in generation diversity (Chen et al., 2025; O'Mahony et al., 2024; Lin et al., 2025) severely constrains the model's exploration space during the subsequent RL phase. To mitigate mode collapse, existing studies have proposed extending beyond pure CE loss by modifying the update policy or incorporating diversity regularization terms. However, their application is often restricted to the RL stage (Wang et al., 2025), or the performance trade-offs incurred during the SFT phase may not be fully recovered in subsequent RL steps.

Through an analysis of mathematical tasks, we identify the token-level exploration space as a critical factor, and this non-uniformity is a critical factor hindering accuracy when blindly encouraging diversity. To this end, we introduce SED-SFT (Selectively Encouraging Diversity in Supervised Fine-Tuning), which incorporates a concise regularization term to modulate the prediction probabilities of target labels, alongside a masking strategy designed to prevent excessive diversity enhancement in regions with limited exploration space. Our research focuses primarily on mathematical reasoning tasks. We evaluated SED-SFT using two prominent base models across eight rigorous mathematical benchmarks. The results demonstrate significant performance gains, most notably an average improvement of 2.06 points on the Llama-3.2-3B-Instruct model. A detailed analysis of the training process confirms that SED-SFT effectively enhances generative diversity during the SFT stage, acting as a vital catalyst for the superior performance observed in the subsequent RL phase.

Our main contributions are summarized as follows:

- We identify and analyze the discrepancies

* Equal contribution.

[1] The code is publicly available at https://github.com/pppa2019/SED-SFT

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 656–663

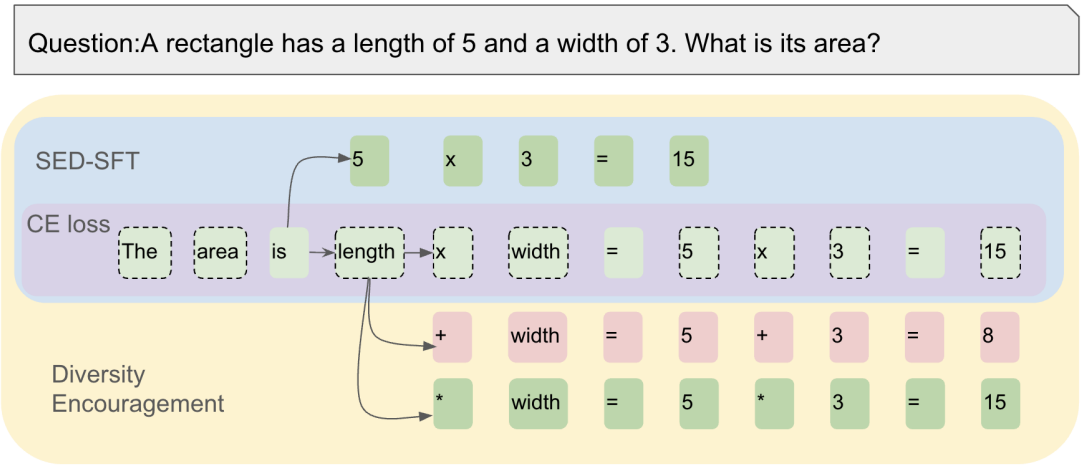

Figure 1: The comparison of Cross-Entropy, pure diversity-encouraging, and SED-SFT. The dashed line token boxes indicate the tokens will be masked in SED-SFT, *i.e.*, the tokens with low exploration space. SED-SFT achieves a balance between accuracy and diversity by avoiding encouraging the masked tokens.

within the token-level exploration space in the diversity encouraging process, thereby justifying the necessity of selective updating.

- We propose SED-SFT, including a concise entropy penalty to foster diversity and a masking strategy to maintain predictive precision.
- Experiments conducted on two mainstream backbones across eight math benchmarks indicate that SED-SFT significantly improves model performance after the RL process, consistently outperforming the baselines.

2 Preliminary Analysis

This section provides a systematic analysis of the token-level exploration space in SFT. We start by formulating the Cross-Entropy objective, then show why preserving diversity is unnecessary for certain tokens, and finally introduce cumulative top-k probability as a practical metric for quantifying exploration space.

2.1 On Cross-Entropy in SFT

SFT trains the model π_θ by minimizing the standard CE loss $\mathcal{L}_{\text{CE}}$:

$$\mathcal{L}_{\text{CE}}(\theta) = -\mathbb{E}_{(x,y^*)\sim\mathcal{D}}\left[\sum_{t=1}^{|y^*|}\log \pi_\theta(y_t^*|x, y_{<t}^*)\right]$$

where y^* is the given label sequence, this loss formulation strongly drives the model policy π_θ to quickly converge along the single correct path y^*, leading to mode collapse and a significant decline in generation diversity (O'Mahony et al., 2024).

2.2 Discrepancies in Token-level Exploration

By viewing the selection of y^* as a binary event, the probabilities of ground-truth labels serve as the determining factors for the magnitude of the model update (Zhang et al., 2025). To mitigate mode collapse, updates should be strategically restricted to positions that potentially contain alternative reasoning branches. We analyze the ground-truth label probabilities through a case study (see the visualization heatmap in Appendix A Figure 2). The results indicate that **tokens with restricted exploration space**—typically comprising fixed conjunctions, structural tokens, or specific vocabulary—exhibit significantly higher prediction confidence compared to the average prediction score. Consequently, to foster diversity and enrich the model's reasoning paths, preserving diversity on such tokens is unnecessary and may even lead to a degradation in accuracy.

2.3 Quantifying the Token Exploration Space

While the analysis in Section 2.2 reveals that a selective strategy during large-scale SFT requires an automated metric to identify tokens with low diversity requirements. To bridge this gap, we seek a quantitative measure that can reflect the exploration space at each position. Specifically, we evaluate the cumulative top-k probability distribution, a metric that is computationally efficient and robust to semantically equivalent token branches (Kuhn et al., 2023). Our analysis of 100 mathematical problems (Figure 3) reveals a trade-off in the choice of k. While the median cumulative probability is

most discriminative at $k = 1$, it offers limited information regarding potential alternative paths. Conversely, as k increases, the metric tends toward uniformity, losing its ability to distinguish restricted tokens from flexible ones. Therefore, we employ $k = 2$ or 3 to characterize the exploration space effectively, and we use this metric to construct the selective mask in the next section.

3 Selective Diversity Encouragement

In this section, we introduce SED-SFT, a simple yet effective training objective designed to encourage the model to maintain diversity at appropriate positions (see Figure 1).

Top-k Masking Strategy M_t To selectively apply the $\mathcal{L}_{\text{DE}}$, we define a binary mask M_t based on the model's prediction distribution at position t. We quantify the token exploration space using the cumulative probability of the Top-k tokens:

$$P_{\text{Top-k}}(t) = \sum_{j \in \mathcal{K}_t} \pi_\theta(y_j | x, y^*_{<t})$$

$$M_t = \mathbf{1}\left[P_{\text{Top-}k}(t) < \tau\right]$$

where $\mathcal{K}_t$ denotes the set of indices for the k tokens with the highest probabilities at step t. To obtain a suitable threshold, we define the masking ratio r and let $\mathcal{P} = \{P_{\text{Top-k}}(t)\}_{t=1}^{T}$ denote the set of cumulative probabilities over all positions in the sampled subset of the training data. We define the threshold τ as the $(1-r)$-th quantile of $\mathcal{P}$:

$$\tau = \text{Quantile}(\mathcal{P}, 1 - r)$$

Diversity Encouraging Function $\mathcal{L}_{\text{DE}}(p)$ The objective of $\mathcal{L}_{\text{DE}}(p)$ is to suppress the probability p of the correct label by pushing it towards 0.5, which corresponds to the point of maximum information entropy. We adopt a concise quadratic function, inspired by CHORD (Zhang et al., 2025), as the penalty term. This function attains its minimum penalty at $p = 0.5$ and maximal penalty at $p = 1.0$ or $p = 0.0$.

$$\mathcal{L}_{\text{DE}}(p) = \left(p - \frac{1}{2}\right)^2$$

Here, $p = \pi_\theta(y^*_t | x, y^*_{<t})$ is the probability assigned to the ground-truth token y^*_t. By minimizing $\mathcal{L}_{\text{DE}}$, p is encouraged to approach 0.5, thereby allocating more probability mass to alternative plausible paths.

Overall Loss $\mathcal{L}_{\text{SED-SFT}}$ The final loss combines the diversity encouraging function and the CE loss. Let λ be the weight of the diversity encouraging term, and all of our experiments use $\lambda = 1$.

$$\mathcal{L}_{\text{SED-SFT}}(\theta) = \sum_{t=1}^{|y^*|} \Big[-\log \pi_\theta(y^*_t \mid x, y^*_{<t}) + \lambda \cdot M_t \cdot \mathcal{L}_{\text{DE}}\big(\pi_\theta(y^*_t \mid x, y^*_{<t})\big) \Big] \tag{1}$$

We next validate SED-SFT under the standard SFT-then-RL pipeline and compare it with representative SFT objectives in Section 1.

4 Experiments

4.1 Experimental Settings

All experiments follow a standard SFT-then-RL pipeline to evaluate whether selectively encouraging diversity during SFT improves downstream RL performance. Unless otherwise specified, we keep all training and evaluation settings identical across methods and vary only the SFT objective.

Backbone Models We conduct experiments on two instruction-tuned backbones: Qwen2.5-Math-7B-Instruct[2] and Llama-3.2-3B-Instruct[3].

Datasets For SFT, we sample 20,000 examples from the Micomind dataset[4]. For RL, we use the Math (Level 1) training split[5].

Training Details For SFT, we follow GEM's training setup, using a learning rate of 2×10^{-5} and DeepSpeed stage-2[6]. For RL, we apply GRPO (Shao et al., 2024) implemented in the Verl framework[7] with batch size 256; other hyperparameters use the default GRPO configuration. We generate RL training samples by sampling each prompt 8 times with Qwen2.5-Math-7B-Instruct and filter out prompts where all samples either fail or succeed (similar to an offline DAPO procedure (Yu et al., 2025)), resulting in 2,069 filtered samples from an initial 5,000.

Compute Devices All SFT/RL training and evaluation are conducted on 8 NVIDIA H20 GPUs.

[2] https://huggingface.co/Qwen/Qwen2.5-Math-7B-Instruct

[3] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[4] https://huggingface.co/micromind

[5] https://huggingface.co/datasets/DigitalLearningGmbH/MATH-lighteval

[6] https://github.com/deepspeedai/DeepSpeed

[7] https://github.com/volcengine/verl

		Avg@8								
		aime24	aime25	amc23	gsm8k	math	gaokao	olympiad	college	average
					Qwen2.5-Math-7B-Instruct					
SFT	CrossEntropy	7.10	14.20	46.90	87.70	66.80	59.50	26.10	38.30	43.33
	GEM	5.00	9.20	35.30	86.80	67.70	56.60	28.00	38.50	40.89
	DFT	17.10	17.50	58.10	95.30	83.90	70.10	44.00	47.90	54.24
	SED-SFT w/o mask	6.20	11.70	40.00	87.40	67.00	59.70	27.70	38.20	42.24
	SED-SFT	6.70	12.10	42.20	87.70	67.50	57.70	26.50	37.90	42.29
RL	CrossEntropy	15.40	16.20	**67.80**	94.50	85.80	72.50	46.50	**49.30**	56.00
	GEM	12.10	16.70	63.70	**96.00**	**87.00**	70.60	48.90	48.70	55.46
	DFT	16.70	18.30	61.60	**96.00**	85.40	70.60	45.80	48.50	55.36
	SED-SFT w/o mask	**20.00**	16.70	67.20	94.90	86.00	71.40	48.30	48.30	56.60
	SED-SFT	18.80	**18.80**	66.20	95.20	86.60	**73.00**	**50.20**	48.80	**57.20**
					Llama-3.2-3B-Instruct					
SFT	CrossEntropy	0.80	1.20	17.20	60.70	34.10	34.50	9.60	19.90	22.25
	GEM	0.00	1.20	10.30	59.20	31.60	33.80	9.90	18.60	20.58
	DFT	2.10	0.80	16.60	69.90	39.50	37.10	12.10	25.30	25.43
	SED-SFT w/o mask	0.00	1.20	12.80	60.10	34.10	33.00	10.70	19.90	21.48
	SED-SFT	0.80	1.20	16.20	61.20	34.30	32.70	9.60	20.30	22.04
RL	CrossEntropy	9.20	0.00	33.40	83.50	56.40	47.80	20.90	34.00	35.65
	GEM	6.50	1.10	27.30	81.00	51.60	44.90	16.90	32.00	32.66
	DFT	2.50	**1.20**	26.60	78.60	49.60	41.00	17.80	30.30	30.95
	SED-SFT w/o mask	**13.60**	0.60	31.50	84.00	54.60	41.00	21.00	**34.50**	35.10
	SED-SFT	11.20	0.40	**38.40**	**85.10**	**57.70**	**53.00**	**22.20**	33.70	**37.71**

Table 1: The main results of Llama-3.2-3B-Instruction and Qwen2.5-Math-7B-Instruct across 8 math datasets. Avg@8 represents the average pass rates over eight sampling iterations for the AIME24/25 and AMC23.

4.2 Evaluation Settings

The evaluation framework follows Qwen-2.5-Math[8]. All of the math benchmarks are widely used in existing works, including: AIME24 (Zhang and Math-AI, 2024), AIME25 (Zhang and Math-AI, 2025), AMC23[9], GSM8K (Cobbe et al., 2021), MATH500 (Hendrycks et al.), GAOKAO-en (Zhang et al., 2023), OlympiadBench (He et al., 2024), and College-MATH [10], and the corresponding dataset name in the evaluation framework is: aime24, aime25, amc23, gsm8k, math, gaokao2023en, olympiadbench, and college_math. We sample 8 times for AIME24/25 and AMC23 (the temperature is 0.7), and sample 1 time for other benchmarks (the temperature is 1.0), and then use the average pass rate as the metric.

4.3 Baselines Selection

We select the following representative objectives as SFT baselines: (1) Cross-Entropy, the most widely used training loss in SFT; (2) GEM (Li et al., 2024), which combines reverseKL and entropy loss in SFT; (3) DFT (Wu et al., 2025), which uses a confidence score to re-weight each token to debias the update. We further include *SED-SFT w/o mask* as an ablation to isolate the effect of the proposed selective masking strategy.

4.4 Main Results

The main results are presented in Table 1. The mathematical task evaluation was conducted on two major base models, yielding the following key findings: (1) On the base models Qwen-2.5-7B-Math and Llama-3.2-3B-Instruct, SED-SFT significantly outperformed the baseline methods. Specifically, compared to the cross-entropy baseline, SED-SFT achieved improvements of 1.20 and 2.06 points, respectively, averaged across eight widely used mathematical benchmarks. (2) DFT and GEM, two representative SFT optimization methods, failed to maintain consistent success. DFT demonstrated a substantial advantage during the SFT phase, significantly surpassing all other methods. However, its approach greatly restricts the model's exploration space, making further improvements in the subsequent RL phase unfeasible. In contrast, GEM encourages overall diversity

[8] https://github.com/QwenLM/Qwen2.5-Math

[9] https://huggingface.co/datasets/math-ai/amc23

[10] https://huggingface.co/datasets/di-zhang-fdu/College_Math_Test

but ignores the token exploration space, resulting in limited adaptability to mathematical tasks. (3) The "w/o mask" in the table represents the ablation study of the SED-SFT method with the selective mask removed. The results demonstrate that the selective mask consistently enhances the final performance of the models across different base models, thereby validating its effectiveness.

5 Analysis

5.1 Sentence-level Diversity Analysis

To assess sentence-level diversity, we employ the Self-BLEU metric for SED-SFT and baseline methods. A lower Self-BLEU score indicates higher diversity in model-generated sentences. As shown in Table 2, both SED-SFT and GEM achieve substantially lower Self-BLEU scores compared to CE and DFT, reflecting enhanced generation diversity.

	Self-BLEU ↓
CE	43.12
DFT	51.26
GEM	38.53
SED-SFT	35.57

Table 2: The Self-BLEU of the generated results of AIME based on Llama-3.2-3B-Instruct.

5.2 Hyperparameter Sensitivity Study

First, we perform parameter sensitivity examination on the ratio r and the k value of top-k in our method, which is shown in Table 3. The results demonstrate that SED-SFT consistently outperforms cross-entropy (CE) when the ratio r exceeds 0.5. Additionally, it is crucial to set the k value greater than 1 to facilitate a more robust evaluation of the token exploration space.

	Average
CE	35.65
SED-SFT($r = 0.2; k = 2$)	34.18
SED-SFT($r = 0.5; k = 2$)	35.73
SED-SFT($r = 0.7; k = 2$)	37.71
SED-SFT($r = 0.8; k = 2$)	35.68
SED-SFT($r = 0.7; k = 1$)	34.68
SED-SFT($r = 0.7; k = 3$)	36.89

Table 3: Sensitivity test on the hyperparameters.

5.3 Robustness across Model Scales

The 3B model possesses weaker foundational capabilities. If diversity is encouraged on tokens with limited exploration space (*e.g.,* format markers, fixed collocations), it is prone to inducing errors and disrupting the reasoning chain; therefore, the mask mechanism is significantly more important for small models. In contrast, the 7B model exhibits greater robustness. Overall, the introduction of the mask mechanism yields more consistent improvements across various dimensions for both the 3B and 7B models.

6 Related Work

Prior work has explored improving exploration efficiency mainly at the RL stage (Wang et al., 2025; Cui et al., 2025; Zhu et al., 2025b). These approaches typically rely on entropy-based control during policy updates, and are therefore not directly applicable to the SFT stage where mode collapse is induced by token-level cross-entropy fitting. To contextualize SED-SFT, we categorize related SFT optimization methods into three streams in Appendix Section B.

7 Conclusion

We propose SED-SFT, a method that utilizes entropy information to enhance supervised fine-tuning by incorporating a simple entropy regularization function and a masking strategy quantifying the token exploration space using cumulative top-k token probability. Across experiments on Llama-3.2-3B-Instruct and Qwen2.5-Math-7B-Instruct and 8 mathematics benchmarks in the SFT-then-RL paradigm, the experimental results demonstrate that SED-SFT consistently enhances the model's generation diversity and consistently improves performance after RL. Meanwhile, according to the analysis model generation result, SED-SFT increases the sentence-level diversity significantly.

Limitations

Limited by computational resources, our experiments are currently restricted to models with a size below 8B. We plan to scale up the model size by utilizing techniques such as LoRA in the future. On the other hand, our experiments are limited to mathematical tasks. The transferability to broader fields such as code generation or universal instruction-following should be discussed in the future.

References

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. 2025. Retaining by doing: The role of on-policy data in mitigating forgetting. *arXiv preprint arXiv:2510.18874*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, and 1 others. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, and 1 others. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Gyuhak Kim, Sumiran Singh Thakur, Su Min Park, Wei Wei, and Yujia Bao. 2025. Sft-go: Supervised fine-tuning with group optimization for large language models. *arXiv preprint arXiv:2506.15021*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. 2024. Preserving diversity in supervised fine-tuning of large language models. *arXiv preprint arXiv:2408.16673*.

Xiaofeng Lin, Hejian Sang, Zhipeng Wang, and Xuezhou Zhang. 2025. Debunk the myth of sft generalization. *arXiv preprint arXiv:2510.00237*.

Rui Ming, Haoyuan Wu, Shoubo Hu, Zhuolun He, and Bei Yu. 2025. One-token rollout: Guiding supervised fine-tuning of llms with policy gradient. *arXiv preprint arXiv:2509.26313*.

Laura O'Mahony, Leo Grinsztajn, Hailey Schoelkopf, and Stella Biderman. 2024. Attributing mode collapse in the fine-tuning of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, volume 2.

Zhiwen Ruan, Yixia Li, He Zhu, Yun Chen, Peng Li, Yang Liu, and Guanhua Chen. 2025. Enhancing large language model reasoning via selective critical token fine-tuning. *arXiv preprint arXiv:2510.10974*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. 2025. On the generalization of sft: A reinforcement learning perspective with reward rectification. *arXiv preprint arXiv:2508.05629*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. 2025. On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. *arXiv preprint arXiv:2508.11408*.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.

Yifan Zhang and Team Math-AI. 2024. American invitational mathematics examination (aime) 2024.

Yifan Zhang and Team Math-AI. 2025. American invitational mathematics examination (aime) 2025.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

He Zhu, Junyou Su, Peng Lai, Ren Ma, Wenjia Zhang, Linyi Yang, and Guanhua Chen. 2025a. Anchored supervised fine-tuning. *arXiv preprint arXiv:2509.23753*.

Wenhong Zhu, Ruobing Xie, Rui Wang, Xingwu Sun, Di Wang, and Pengfei Liu. 2025b. Proximal supervised fine-tuning. *arXiv preprint arXiv:2508.17784*.

A Prediction Probability Heatmap

As shown in Figure 2, we visualize the probability distribution of labels at each position using heatmaps to demonstrate that tokens with exceptionally high probabilities typically correspond to those with limited exploration space.

B Heatmap on the statistic of the cumulative probability of top-k

We sample 100 math task examples from the OpenR1_Math [11] dataset and obtain the cumulative probability for different k values across four representative backbone models: Llama-3.2-3B-Instruct, Qwen2.5-Math-1.5B-Instruct, Qwen2.5-Math-7B-Instruct, and Qwen3-8B. As shown in Figure 3. As illustrated in Figure3, the results reveal significant divergence in cumulative probability distributions across different models and samples.

C Detailed Related Work

We contextualize our approach by discussing existing literature across three main paradigms: RL policy integration, diversity modeling, and selective gradient updating.

RL Policy Integration DFT (Wu et al., 2025) incorporates RL policy-update ideas into SFT by preventing overly large gradients during model updates to preserve the model's exploration capacity. ASFT (Zhu et al., 2025a) further improves the performance based on DFT by solving the bias shifting problem. However, ASFT has a much higher computational cost for inducing a reference model. While this line adjusts the update direction, it does not fundamentally increase diversity and can reduce exploration space, thereby limiting subsequent RL improvement.

Diversity Modeling GEM (Li et al., 2024) explicitly integrates diversity into the objective function to directly encourage varied generation. However, this approach overlooks the distinct nature of different tokens, which can potentially lead to degradation in areas where accuracy should be prioritized. Consequently, they lack robustness in tasks with high accuracy requirements (*e.g.*, mathematics), leading to decreased model performance.

Selective Gradient Updating Selective updating strategies, such as CTF (Ruan et al., 2025) require pre-determining the important update locations, which typically demands favorable model characteristics, high-cost algorithms, or prior knowledge. This results in extremely high inconsistency and poor generalizability for token selection algorithms in general models (Ming et al., 2025; Kim et al., 2025).

[11] https://huggingface.co/datasets/open-r1/OpenR1-Math-220k

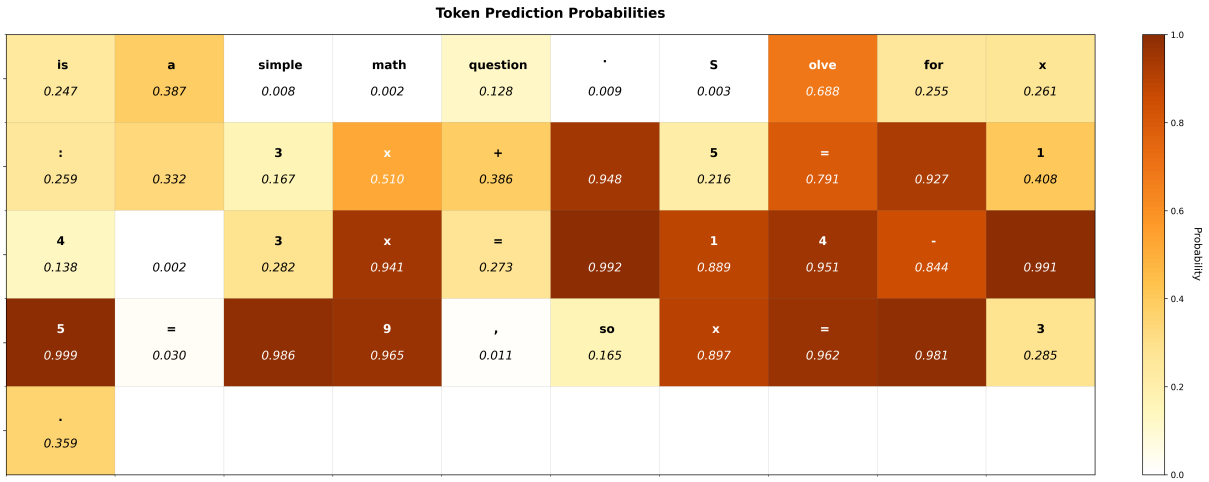

Figure 2: The heatmap for the probability of the labels on each position. The case is a simple math problem and evaluated on Qwen-0.5B

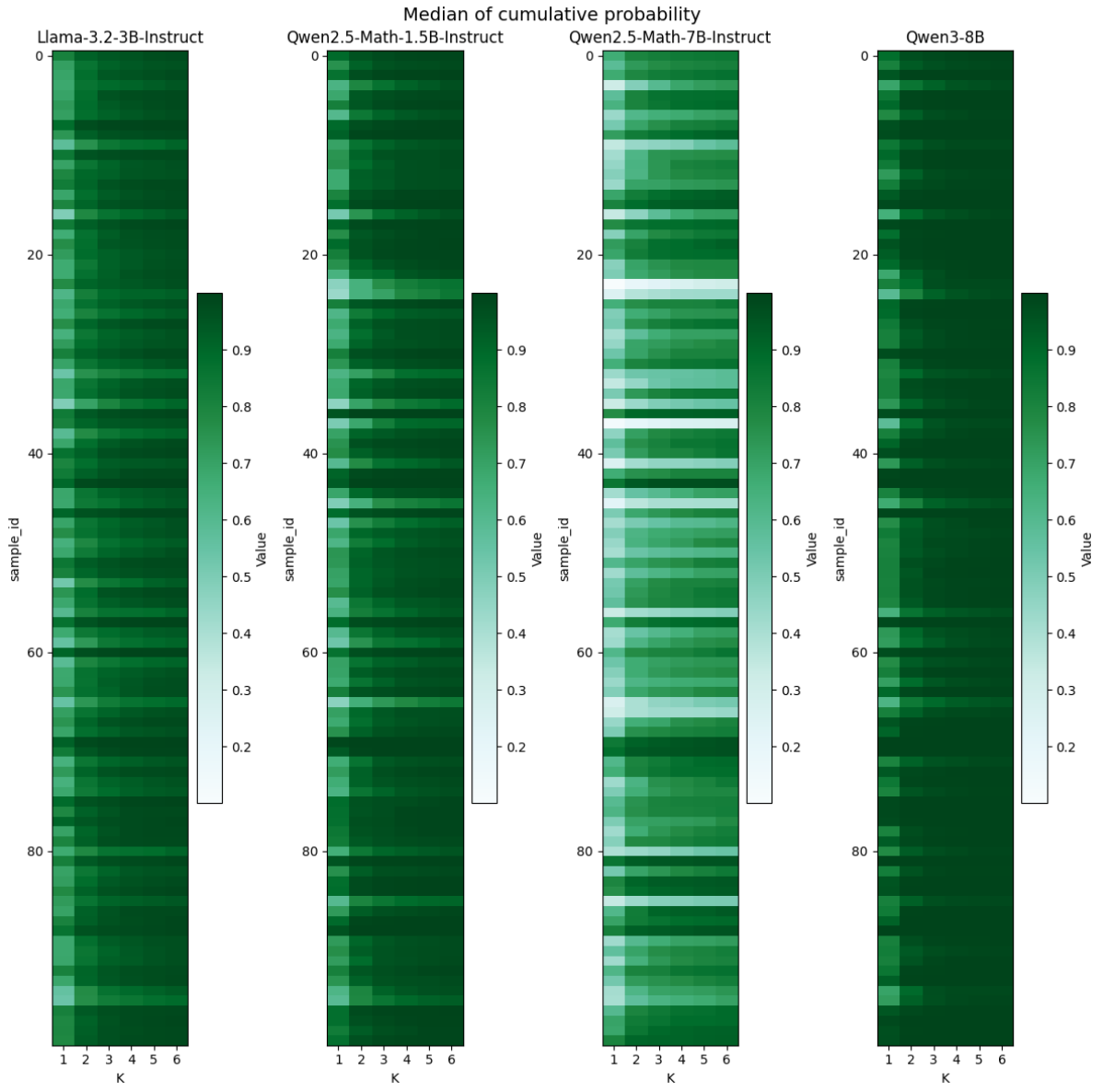

Figure 3: The visualization of the cumulative probability with different k values in Top-k.