Offline Exploration-Aware Fine-Tuning for Long-Chain Mathematical Reasoning

Anonymous ACL submission

Abstract

Through encouraging self-exploration, reinforcement learning from verifiable rewards (RLVR) has significantly advanced the mathematical reasoning capabilities of large language models. As the starting point for RLVR, the capacity of supervised fine-tuning (SFT) to memorize new chain-of-thought trajectories provides a crucial initialization that shapes the subsequent exploration landscape. However, existing research primarily focuses on facilitating exploration during RLVR training, leaving exploration-aware SFT underexplored. To bridge this gap, we propose **O**ffline e**X**ploration-**A**ware (OXA) fine-tuning. Specifically, OXA optimizes two objectives: promoting low-confidence verified teacher-distillation data to internalize previously uncaptured reasoning patterns, and suppressing high-confidence incorrect self-distillation data to redistribute probability mass of incorrect patterns toward potentially correct candidates. Experimental results across 6 benchmarks show that OXA consistently improves mathematical reasoning performance, especially achieving an average gain of $+6$ Pass@1 and $+5$ Pass@k points compared to conventional SFT on the `Qwen2.5-1.5B-Math`. Crucially, OXA elevates initial policy entropy, and performance gains persist throughout extensive RLVR training, demonstrating the long-term value of OXA.

1 Introduction

The mathematical reasoning capabilities of large language models (LLMs) have witnessed a breakthrough by scaling up inference computation for long-chain reasoning. Building upon pre-trained LLMs, a common strategy to achieve state-of-the-art performance involves a two-stage training pipeline (DeepSeek-AI, 2025; Yang et al., 2025a): (1) Supervised fine-tuning (SFT), which distills knowledge from teacher models to activate initial reasoning capabilities and learn long-chain output formats; and (2) Reinforcement learning from verifiable rewards (RLVR), which further boosts performance by encouraging self-exploration and learning from model-generated samples.

In reinforcement learning, maintaining sufficient policy entropy to prevent premature convergence and encourage exploration is fundamental (Williams and Peng, 1991; Williams, 1992; Eysenbach and Levine, 2021). In the context of RLVR, thorough exploration is particularly critical, as it allows the model to discover diverse reasoning paths and unlock greater potential. To this end, existing research has attempted to manipulate policy entropy through objective-level regularizations (Zhang et al., 2025b; Cheng et al., 2025; Jiang et al., 2025), fine-grained update and sampling controls (Liao et al., 2025; Cui et al., 2025), and semantic-level abstractions (Cao et al., 2025). However, these efforts primarily focus on facilitating exploration during the RLVR process, overlooking the role of the SFT stage. As the starting point for RLVR, SFT provides a crucial initialization that shapes the subsequent exploration landscape.

Moreover, recent studies reveal that while RLVR excels at optimizing known paths, it struggles to expand the model's fundamental reasoning space. In contrast, SFT is highly effective at enabling models to internalize new reasoning pathways (Yue et al., 2025; Kim et al., 2025; Chu et al., 2025). This suggests that by enriching the model's exploration space with diverse reasoning pathways, intuitively, SFT can facilitate exploration in the RLVR process. Despite this potential, current long-chain reasoning SFT research focuses exclusively on activating reasoning capabilities (Liu et al., 2025; Guha et al., 2025) or data pruning for efficiency (Muennighoff et al., 2025; Yang et al., 2025b). This work addresses the question: *How can we train exploration-engaged models for RLVR via fine-tuning?*

We envision that an ideal initialization for RLVR should exhibit two key characteristics: *superior*

initial reasoning accuracy and *high initial policy entropy*. To achieve this, we propose **O**ffline e**X**ploration-**A**ware (OXA) fine-tuning, an algorithm designed to train on strategically selected offline reasoning trajectories. To counteract the entropy collapse illustrated in Figure 1, OXA reinforces low-probability trajectories while weakening high-probability ones to preserve the smoothness of the predictive distribution. Specifically, OXA optimizes two objectives: *promoting low-confidence verified teacher-distillation data* and *suppressing high-confidence incorrect self-distillation data*. The former internalizes previously uncaptured reasoning trajectories, while the latter redistributes probability mass of incorrect paths toward potentially correct candidates. Since these objectives are decoupled, we introduce two variants: a base version of OXA utilizing only the first objective and the full OXA framework. While the base version accounts for the primary performance gains, the full framework synergizes superior performance with robust exploration potential.

We evaluate OXA by applying the SFT-then-RLVR paradigm to 4 LLMs ranging from 1.5B to 7B parameters. Experimental results across 6 typical mathematical benchmarks show that OXA consistently enhances reasoning capabilities. Notably, it achieves an average improvement of $+6$ Pass@1 and $+5$ Pass@k points compared to conventional SFT on the 1.5B LLM. Comprehensive analysis further demonstrates that OXA not only improves performance on challenging problems but also significantly expands the reasoning output space, achieving high initial policy entropy. Crucially, these gains persist throughout extensive RLVR training and are orthogonal to existing RLVR-enhancement methods, yielding consistent additive improvements.

2 Related Work

Training long-chain reasoning LLMs. There are primarily three trajectories for training long-chain reasoning LLMs. Initial efforts face the scarcity of supervised data with annotated reasoning steps. They leverage RLVR to directly train pre-trained LLMs to explore self-discovered reasoning paths toward correct answers (DeepSeek-AI, 2025; Team et al., 2025). This base-model training process is termed Zero-RL. However, these models often suffer from poor readability and lower performance ceilings (DeepSeek-AI, 2025). A second paradigm involves two stages: first distilling knowledge from teacher models (e.g., Zero-RL models) into pre-trained LLMs via SFT for a cold start, then followed by RLVR. Other works extend this by integrating supervised signals from teachers into RLVR to learn superior reasoning trajectories (Yan et al., 2025; Xu et al., 2025; Lv et al., 2025; Zhang et al., 2025a). In this work, we follow the SFT-then-RLVR paradigm, widely validated by various open-source LLMs (DeepSeek-AI, 2025; Yang et al., 2025a; Xia et al., 2025).

Analysis of SFT and RLVR. Although both SFT and RLVR can train long-chain reasoning models, recent works reveal that their learning mechanisms differ. RLVR exhibits a generalization pattern: constrained by self-generated training samples, it improves accuracy by increasing the probability of correct answers, but fails to expand capability by sampling correct answers outside the model's output space (Yue et al., 2025; Mu et al., 2025). In contrast, SFT tends to memorize training data. By introducing new knowledge during distillation, SFT can enhance the model's capabilities on difficult problems (Kim et al., 2025; Chu et al., 2025).

Policy entropy in RLVR. Rooted in information theory, entropy provides a principled mechanism to manage the exploitation-exploration trade-off. Higher policy entropy indicates that the model is more likely to explore diverse reasoning paths, thereby enhancing performance. To maintain high entropy during early training, various efforts have been made, including restricting the update of high-covariance tokens (Cui et al., 2025), adjusting entropy regularization coefficients (Zhang et al., 2025b), tuning rollout temperature (Liao et al., 2025), adding extra entropy terms to advantage calculations (Cheng et al., 2025), leveraging cumulative entropy regulation (Jiang et al., 2025), and elevating entropy control from the token to the semantic level (Cao et al., 2025). However, previous works predominantly manipulate RLVR training dynamics but overlook SFT, leaving an open question: *How can we encourage models to explore more reasoning trajectories via SFT?*

Long-chain reasoning SFT. Recent SFT research targeting long-chain reasoning generally follows two directions. The first focuses on eliciting reasoning capabilities in pretrained LLMs. Some studies expand reasoning paths by multi-sample distillation from teacher models (Liu et al., 2025;

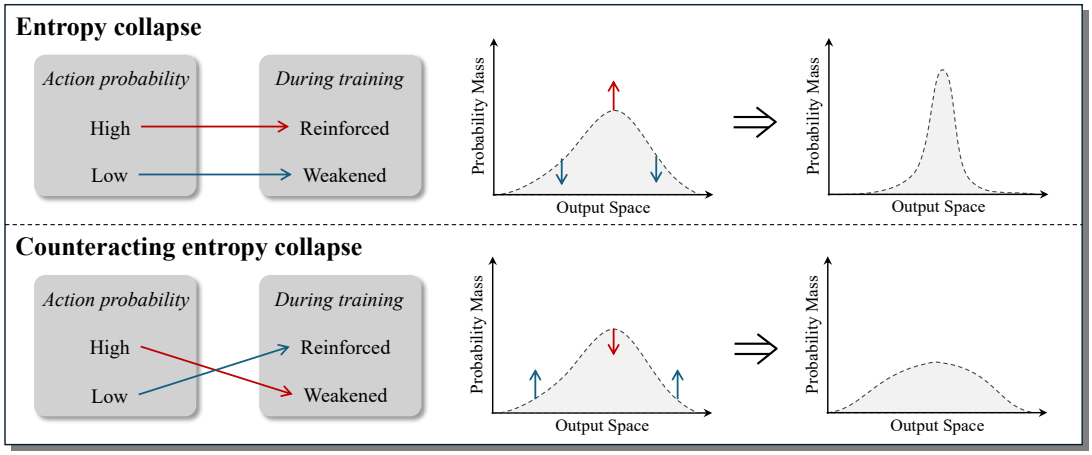

Figure 1: Conceptual illustration of entropy collapse versus counteracting entropy collapse.

Guha et al., 2025), while others employ curriculum learning or decompose complex structures (Wen et al., 2025; An et al., 2025; Luo et al., 2025). The second direction focuses on data pruning to enhance training efficiency (Muennighoff et al., 2025; Yang et al., 2025b). While our method also involves data selection, it is distinct in its objective: rather than optimizing for training efficiency, OXA aims to cultivate exploration-engaged models that provide a superior initialization for subsequent RLVR.

3 Methodology

In the SFT-then-RLVR training paradigm, SFT models serve as the critical foundation for subsequent reinforcement learning. We hypothesize that an ideal SFT model should exhibit superior initial reasoning capabilities while simultaneously fostering exploration-engaged behavior to unlock greater RL potential. Specifically, we aim to achieve two objectives:

- **Achieving higher initial performance:** Providing a robust backbone that maintains or amplifies the performance superiority established during SFT throughout the RLVR process.
- **Maintaining higher policy entropy:** Broadening the reasoning output space and facilitating sampling diverse reasoning trajectories.

To this end, we propose **O**ffline e**X**ploration-**A**ware **(OXA)** fine-tuning, an algorithm designed to train on strategically selected offline reasoning trajectories. Specifically, inspired by counteracting entropy collapse, OXA optimizes the model by promoting low-confidence correct answers and suppressing high-confidence errors. The former increases the likelihood of generating previously uncaptured reasoning trajectories, particularly those near the distribution boundaries, while the latter redistributes probability mass of incorrect reasoning paths toward other potentially correct candidates. Ultimately, OXA yields a model that combines enhanced performance with high initial policy entropy.

As a purely SFT-based approach, OXA offers two distinct advantages: First, it enhances the exploration capability without changing the RLVR framework, thereby preserving training stability while delivering consistent performance gains. Second, empirical results demonstrate that OXA is orthogonal to RLVR-enhancement methods, providing additive improvements.

3.1 Dissecting Entropy Dynamics

Policy entropy, quantifying the smoothness of the predictive distribution of model π_θ, is defined as:

$$\mathcal{H}(\pi_\theta) = -\sum_{i=1}^{|\mathcal{V}|} p_i \log p_i, \tag{1}$$

where p_i represents the probability of the i-th token in the vocabulary $\mathcal{V}$ of size $|\mathcal{V}|$. In the context of training long-chain mathematical reasoning LLMs, entropy serves as a direct proxy for the diversity of reasoning paths the model can sample. A higher policy entropy indicates a more uniform distribution of probability mass, enabling the model to explore more candidate outputs. Conversely, a lower entropy characterizes a sharp distribution where the model's confidence is excessively concentrated on a limited set of tokens, thereby restricting its exploration space and limiting the variety of generated reasoning trajectories.

Entropy collapse. When a model is trained to convergence, the entropy generally decreases. As illustrated in the upper one of Figure 1, this phenomenon stems from the model being highly aligned with the empirical distribution of the training data, including reinforcing the distribution peaks while suppressing the troughs.

Counteracting entropy collapse. To mitigate entropy collapse, a straightforward strategy is to inversely influence the distribution dynamics: promoting the probability mass at the distribution troughs while suppressing over-confident peaks, which is depicted in the lower one of Figure 1.

3.2 Offline Exploration-Aware Fine-tuning

A higher policy entropy is preferred for exploration-engaged models. However, directly reinforcing the probability at the trough and weakening that at the peak can severely destabilize the model. OXA provides a solution that selectively promotes low-confidence truths and suppresses high-confidence errors. All reasoning instruction data for this process is curated offline.

3.2.1 Promote Low-Confidence Truths

This objective aims to reinforce the model's probability mass in low-confidence regions by training on verified teacher-distilled data via the maximum likelihood estimation (MLE) criterion. Through this process, the model internalizes previously uncaptured reasoning trajectories, thereby effectively expanding its reasoning space.

Data. Based on the teacher-distilled dataset, we first employ the rule-based verifier used during RLVR training to filter the data, retaining only correct reasoning paths. We then use perplexity (PPL) to quantify the model's confidence in a specific reasoning route. For a reasoning trajectory $S = \{s_1, s_2, \ldots, s_K\}$, the PPL is defined as:

$$\mathrm{PPL}(S) = \exp\left(-\frac{1}{K}\sum_{t=1}^{K}\log p(s_t \mid s_{<t})\right), \quad (2)$$

where K denotes the sequence length and $p(s_t \mid s_{<t})$ is the conditional probability assigned by the model π_θ. A higher PPL indicates that the model is less confident in the trajectory, signifying a hard-to-sample reasoning path, while a lower PPL suggests the opposite.

To prevent the training set from being dominated by excessively difficult learning samples, we design a Gaussian-guided PPL sampling algorithm (Algorithm 1). This algorithm samples data according to a predefined Gaussian PPL distribution consisting of three key stages: binning, target distribution setup, and length-prioritized sampling. It allows us to explicitly control the PPL distribution centered at μ with a dispersion σ, while enforcing maximum responses per query d. Specifically, an increase in μ shifts the selection toward higher-PPL reasoning paths, while σ modulates the sampling density for data points deviating from the central perplexity. Furthermore, within the same PPL bin, we prioritize longer responses to enhance the model's capability in generating complex, multi-step reasoning chains. Hyperparameter details for this sampling process are provided in Appendix A.2.

Algorithm 1 Gaussian-Guided PPL Sampling

Input: Dataset $\mathcal{D} = \{(q, r, p, l)\}$ containing query, response, PPL, and length; Hyperparameters μ, σ, total size N, maximum responses per query d, PPL range $[p_{\min}, p_{\max}]$ and bin width w.
Output: Selected subset $\mathcal{R}$.

Step 1: Binning
Define $M = \lceil (p_{\max} - p_{\min})/w \rceil$ bins.
Partition $\mathcal{D}$ into bins $\mathcal{B}_1, \ldots, \mathcal{B}_M$ based on PPL p.
Discard samples where $p \notin [p_{\min}, p_{\max}]$.
Step 2: Target Distribution Setup
for each bin $i \in \{1, \ldots, M\}$ **do**
  Let x_i be the center PPL of bin i.
  Compute density $d_i = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_i-\mu}{\sigma})^2}$.
end for
Normalize counts: $T_i \leftarrow \lfloor N \cdot (d_i / \sum_j d_j) \rfloor$.
Step 3: Length-Prioritized Sampling
for each bin $i \in \{1, \ldots, M\}$ **do**
  Sort $\mathcal{B}_i$ by length l in descending order.
  Initialize bin counter $c_i \leftarrow 0$.
  for each candidate $(q, r, p, l) \in \mathcal{B}_i$ **do**
    if $c_i < T_i$ **and** $Count(q) < d$ **then**
      $\mathcal{S} \leftarrow \mathcal{S} \cup \{(q, r)\}$.
      $Count(q) \leftarrow Count(q) + 1$, $c_i \leftarrow c_i + 1$.
    end if
  end for
end for
return $\mathcal{R}$

Training. Given a batch of M reasoning trajectories $\mathcal{B}_{\text{MLE}} = \{S_1, S_2, \ldots, S_M\}$ (each with K_S tokens) selected for promotion, we adopt the MLE objective by minimizing the cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{M \cdot K_S}\sum_{S \in \mathcal{B}_{\text{MLE}}}\sum_{t=1}^{K_S}\log p(s_t \mid s_{<t}). \quad (3)$$

3.2.2 Suppress High-Confidence Errors

The second objective of OXA focuses on weakening the probability mass at erroneous peaks by suppressing high-confidence but incorrect reasoning

trajectories via the unlikelihood loss. This process redistributes the probability mass from incorrect paths toward potentially correct alternatives.

Data. Since the rollouts from a pre-trained LLM often suffer from low quality, we first train an instruction-following model using a small set of teacher-distillation data, then use it to sample reasoning trajectories. After verifying trajectories, we calculate the PPL of incorrect ones using the pre-trained LLM to assess confidence. Subsequently, we select the samples with the lowest PPL that fail the verification for suppression.

Training. Given a batch of N reasoning trajectories $\mathcal{B}_{\text{UL}} = \{S_1, S_2, \ldots, S_N\}$ identified as high-confidence errors, we apply the token-level unlikelihood loss (Welleck et al., 2020):

$$\mathcal{L}_{\text{UL}} = -\frac{1}{N \cdot K_S} \sum_{S \in \mathcal{B}_{\text{UL}}} \sum_{t=1}^{K_S} \log\left(1 - p(s_t \mid s_{<t})\right). \tag{4}$$

3.2.3 Global Training Objective

Combining these components, the final OXA objective integrates both losses. We introduce a hyperparameter α to weight the unlikelihood loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha \cdot \mathcal{L}_{\text{UL}}. \tag{5}$$

In practice, α is kept small to prevent excessive gradient magnitudes that could destabilize training. A theoretical analysis demonstrating how a large α leads to vanishing or exploding gradients is provided in Appendix A.1.

We evaluate two variants of our framework: a baseline version using only the first objective (OXA_{MLE}) and the complete framework (OXA_{Full}).

4 Experiments

In this section, we demonstrate that OXA is capable of evolving pre-trained LLMs into exploration-engaged RLVR starting points, possessing superior mathematical reasoning capabilities to unlock further RL potential.

- In Section 4.2, our main experiment shows that OXA fulfills two objectives: improving initial performance and maintaining higher policy entropy.
- In Section 4.3, we validate the effectiveness of OXA when generalizing to two additional LLMs and scaling the training data.
- In Section 4.4, we present extensive ablation studies, including validating each component and hyperparameter, analyzing the impact of long data, assessing generalization to out-of-distribution reasoning tasks, and demonstrating orthogonality to other methods.

4.1 Experiment Setups

Model. We select several pre-trained LLMs as the training start points. Our main experiment is based on `Qwen2.5-1.5B-Math` and `Qwen2.5-7B-Math`, hereafter referred to as the 1.5B and 7B models, respectively. Our main experiment follows the full SFT-then-RLVR paradigm by applying different SFT methods while maintaining the unified RLVR settings. In our model generalization experiment, we also validate OXA on `LLaMA3.2-3B-Base` and `Qwen3-1.7B-Base`.

Baselines. We choose two baselines. The first is conventional SFT, which uses teacher distillation data as supervised signals to fine-tune pre-trained LLMs. The second is low-PPL preferred SFT (SFT_{LP}), which fine-tunes models using distillation data with the lowest PPL, serving as a contrast to our approach. To ensure fairness, as OXA utilizes $50,000$ teacher distillation samples, we select an equal amount of data randomly for conventional SFT and based on low-PPL preference for SFT_{LP}. Furthermore, in Section 4.4, we compare OXA with conventional SFT using the full dataset with 2.6 million samples.

SFT Details. We use `AceReason-1.1-SFT`[1], a pollution-free dataset containing 2.6 million unverified DeepSeek-R1 distilled mathematical samples. After tracing back original answers, we use `math-verify`[2] to filter out incorrect reasoning paths, leaving nearly 2 million samples. We then apply our sampling algorithm to select $50,000$ high-PPL correct samples and $50,000$ low-PPL incorrect ones. We separately evaluate OXA_{MLE} and OXA_{Full}. All of the instruction datasets maintain a one-query-to-one-response ratio. Figure 2 (a) illustrates the distribution of sequence lengths and PPL for the 7B model's training data. During fine-tuning, we use a batch size of 128 and a UL loss weight of 10^{-4} for 6 epochs, with learning rates of 2.5×10^{-4} (1.5B) and 5×10^{-5} (7B). See Appendix A.3 for more details.

[1] https://huggingface.co/datasets/nvidia/AceReason-1.1-SFT

[2] https://github.com/huggingface/Math-Verify

Model	AIME24		AIME25		BRUMO25		CMIMC25		HMMT25		Minerva		Avg. Perf.	
	Pass@1	Pass@128	Pass@1	Pass@128	Pass@1	Pass@128	Pass@1	Pass@128	Pass@1	Pass@128	Pass@1	Pass@64	Pass@1	Pass@k
Qwen2.5-1.5B-Math														
Base	9.4	53.3	4.5	36.7	11.5	40.0	2.4	35.0	0.5	26.7	11.9	62.9	6.7	42.4
SFT_{LP}	20.5	76.7	20.5	53.3	29.8	73.3	9.8	50.0	7.1	40.0	20.7	61.4	18.1	59.1
SFT	23.2	80.0	23.8	60.0	29.0	80.0	11.0	52.5	11.6	50.0	22.3	61.4	20.2	64.0
OXA_{MLE}	35.0	**83.3**	**27.1**	**66.7**	**36.1**	**83.3**	14.9	57.5	**18.5**	**60.0**	**25.7**	63.2	**26.2**	69.0
OXA_{Full}	**35.4**	80.0	26.7	**66.7**	34.7	**83.3**	**15.9**	**62.5**	18.2	**60.0**	22.8	**64.0**	25.6	**69.4**
SFT_{LP}†	22.3	70.0	22.6	60.0	29.4	76.7	10.7	45.0	9.5	46.7	22.0	58.5	19.4	59.5
SFT†	27.2	76.7	25.4	60.0	34.6	76.7	11.9	45.0	13.8	56.7	23.3	62.5	22.7	62.9
OXA_{MLE}†	**40.1**	**83.3**	27.9	**63.3**	39.1	**83.3**	16.4	**60.0**	**19.7**	53.3	27.1	**65.8**	28.4	68.2
OXA_{Full}†	39.0	**83.3**	**29.3**	**63.3**	**41.8**	76.7	**17.4**	57.5	19.1	**66.7**	**27.6**	65.1	**29.0**	**68.8**
Qwen2.5-7B-Math														
Base	16.0	63.3	7.3	43.3	9.8	53.3	2.4	35.0	0.6	16.7	13.2	64.0	8.2	45.9
SFT_{LP}	38.4	83.3	29.0	63.3	41.2	86.7	20.5	60.0	19.3	63.3	34.2	62.1	30.4	69.8
SFT	47.5	86.7	34.2	80.0	48.0	83.3	26.8	77.5	24.4	66.7	33.3	65.4	35.7	76.6
OXA_{MLE}	**54.5**	**90.0**	**39.3**	**83.3**	**50.9**	**90.0**	**26.6**	75.0	**24.5**	**80.0**	**37.3**	**65.4**	**38.8**	**80.6**
OXA_{Full}	50.2	86.7	36.7	**83.3**	**50.9**	**90.0**	24.8	**80.0**	23.1	70.0	36.3	64.7	37.0	79.1
SFT_{LP}†	42.0	76.7	31.6	66.7	46.1	80.0	23.3	55.0	21.3	56.7	36.7	63.2	33.5	66.4
SFT†	50.9	**90.0**	35.1	80.0	51.6	83.3	30.3	**67.5**	23.6	70.0	35.1	64.0	37.8	75.8
OXA_{MLE}†	57.9	83.3	**42.0**	**93.3**	54.3	**90.0**	28.3	65.0	26.5	63.3	38.7	64.3	**41.3**	**76.6**
OXA_{Full}†	**58.9**	83.3	40.8	80.0	**54.4**	**90.0**	**28.4**	60.0	**26.6**	**66.7**	**38.9**	**66.9**	**41.3**	74.5

Table 1: Performance comparison of fine-tuning methods and their corresponding RLVR stages. "Base" denotes the pre-trained LLMs, and †indicates models after RLVR training. Pass@1 scores are averaged over 128 samples, except for Minerva which is averaged over 64 samples. Best results within each SFT/RLVR group are bolded.

RLVR Details. Our training dataset is a subset of DeepScaleR-40K[3]. For RL training, we use a maximum output length of $16,384$, 8 rollouts per prompt, a batch size of 64, and a decoding temperature of 0.85. The learning rate is set to 2×10^{-6}. We train the 1.5B and 7B models for $1,600$ and $1,200$ update steps, respectively. For each RL experiment, we report results from the checkpoint achieving the peak score on the AIME24 benchmark. See Appendix A.3 for other details.

Evaluation. We comprehensively evaluate on six mathematical benchmarks: AIME24, AIME25, BRUMO25 (Balunovic et al., 2025), CMIMC25 (Balunovic et al., 2025), HMMT25 (Balunovic et al., 2025), and Minerva (Lewkowycz et al., 2022). We report Pass@1 and Pass@k scores to assess reasoning capabilities, where the latter represents the model's potential to solve problems. To ensure stability, Pass@1 is averaged over k samples. We set $k = 64$ for Minerva and $k = 128$ for all other datasets. By default, we generate from the models using a temperature of 0.6, a Top-p value of 0.95, and a maximum output length of $32,768$ tokens.

4.2 Main Results

Results of fine-tuning. We evaluate 1.5B and 7B LLMs across four configurations: standard SFT, low-PPL SFT (the converse of OXA), and OXA in its MLE-only variant and its complete form (MLE + UL Objective), denoted as SFT, SFT_{LP}, OXA_{MLE}, and OXA_{Full}, respectively. As summarized in Tables 1 and 2, several key observations emerge: (1) Both OXA variants significantly outperform the baselines; notably, the OXA fine-tuned 1.5B model achieves average improvements of $+6$ Pass@1 points and $+5$ Pass@k points over conventional SFT. These gains suggest that OXA not only enhances base reasoning performance but also augments the model's capability to solve challenging problems. (2) SFT_{LP} consistently underperforms relative to SFT and lags significantly behind OXA_{MLE}, highlighting the effectiveness of internalization low-confidence samples. (3) While OXA_{Full} yields a slightly lower Pass@1 score than OXA_{MLE}, its Pass@k performance remains competitive, indicating robust solution diversity. (4) Despite a selection bias toward longer reasoning trajectories, OXA does not substantially alter the average response length of the model.

Model	SFT_{LP}	SFT	OXA_{MLE}	OXA_{Full}
1.5B	15205	15054	15809	15700
7B	13477	12388	12677	13002

Table 2: Average output lengths of fine-tuned Qwen2.5-1.5B/7B on the AIME24 benchmark.

Results of reinforcement learning. Subsequently, we perform extensive RLVR training on the fine-tuned LLMs, spanning $1,600$ update steps for the 1.5B models and $1,200$ steps for the 7B models. Based on the results marked with †in Table 1, we observe that: (1) The performance gains achieved by OXA persist throughout the RLVR pro-

[3] https://huggingface.co/datasets/agentica-org/DeepScaleR-Preview-Dataset

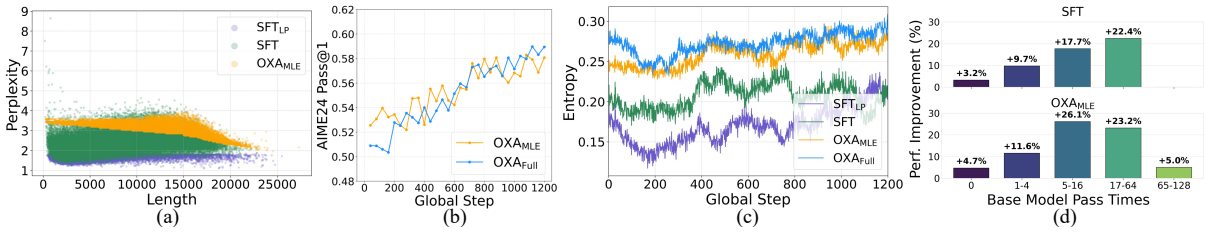

Figure 2: (a) Sequence length and PPL distributions of the 7B model's training data. (b) Performance trajectories of the 7B model fine-tuned with OXA_{MLE} and OXA_{Full} on AIME24. (c) Policy entropy dynamics during RLVR training for various fine-tuning methods based on the 7B model. (d) Performance gains of SFT and OXA_{MLE} on the 1.5B model and the Minerva benchmark, grouped by the base model's pass counts.

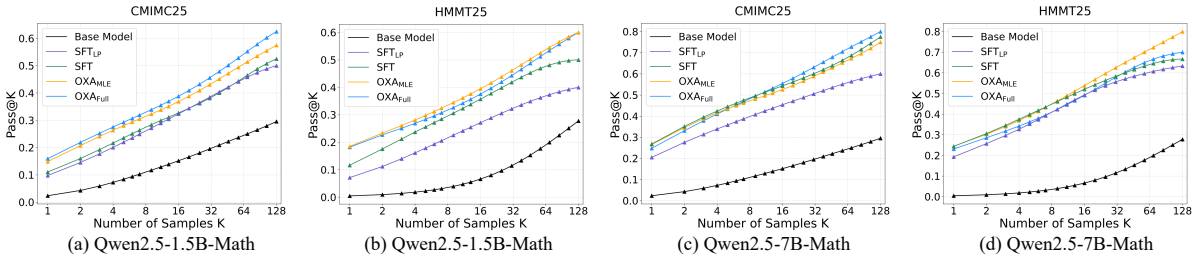

Figure 3: Pass@k curves of 1.5B and 7B models across different fine-tuning methods on CMIMC25 and HMMT25.

cess, yielding more robust reasoning models upon completion of the full SFT-then-RLVR pipeline. (2) Compared to OXA_{MLE}, OXA_{Full} achieves superior or competitive performance after RLVR training. Specifically, Figure 2 (b) illustrates that OXA_{Full} exhibits a more rapid performance ascent during the RL process, ultimately leading to higher scores.

Dynamics of policy entropy. Figure 2 (c) records the policy entropy dynamics for the 7B model during RLVR training. The results indicate that, compared to SFT_{LP} and conventional SFT, OXA models—particularly OXA_{Full}—sustain higher entropy levels during the initial training phase. This validates the effectiveness of our approach in facilitating the sampling of diverse reasoning trajectories.

Frontiers of reasoning potential. To understand the distinct impact of OXA compared to other SFT methods, we categorize Minerva's problems by difficulty based on the pre-trained LLM's pass counts across 128 rollouts. Figure 2 (d) shows that OXA outperforms SFT across all difficulty levels, particularly in the range of 5 to 16, where more challenging problems reside, yielding significant gains. Moreover, Pass@k results in Figure 3 validate that OXA expands the model's reasoning potential. These results consistently demonstrate that OXA increases the likelihood of generating previously uncaptured low-probability reasoning paths, thus enabling the model to solve harder problems.

4.3 Scaling Analysis

Model Generalization. Beyond the Qwen2.5 series, we further evaluate the efficacy of OXA on `LLaMA3.2-3B` and `Qwen3-1.7B` models. Notably, LLaMA3.2 serves as a representative of models that have not undergone extensive pre-training on mathematics corpora. Figure 4 (a) and (b) illustrate the average performance across the 6 mathematical benchmarks. Our results demonstrate that OXA consistently achieves the best results with substantial performance margins. This is particularly evident in the LLaMA3.2 model, where OXA outperforms vanilla SFT by nearly $+6$ Pass@1 points and $+10$ Pass@k points. These findings suggest that OXA yields more pronounced improvements for models lacking mathematics pre-training.

Data Scaling. We further investigate the scalability of OXA by increasing the training data size from $50,000$ to $150,000$ samples. We report the performance of SFT, OXA_{MLE}, and `ARN-1.1-SFT`—a strong baseline fine-tuned on the full `AceReason-1.1-SFT` dataset. `ARN-1.1-SFT` is trained by fine-tuning `Qwen2.5-7B-Math` with 2.6 million mathematical and 1.3 million code reasoning trajectories (Liu et al., 2025). As illustrated in Figure 4 (c), scaling the training data significantly enhances the performance of OXA, which consistently maintains a substantial lead over the SFT baseline. Notably, with only $150,000$ samples, OXA achieves performance nearly on par with

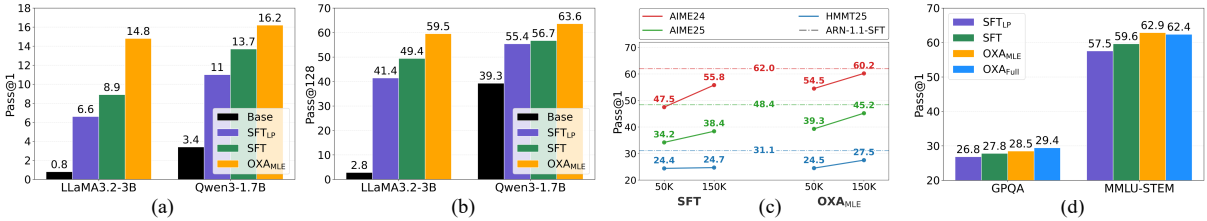

Figure 4: (a)-(b) Average performance across 6 mathematical benchmarks for `LLaMA3.2-3B` and `Qwen3-1.7B` under various fine-tuning methods. (c) Scalability of SFT and OXA_{MLE} as training data increases, compared against ARN-1.1-SFT—a model fine-tuned on millions of samples using the same backbone. (d) Generalization performance of 1.5B variants on out-of-domain benchmarks, including GPQA and MMLU-STEM.

Configuration		AIME24	AIME25
Learning rate (1.5B)	3.0e-4	**23.8**	23.1
	2.5e-4	23.2	**23.8**
	1.0e-4	22.8	23.1
Learning rate (7B)	1.0e-4	45.9	**35.4**
	5.0e-5	**47.5**	34.2
	2.0e-5	45.5	32.7
UL loss weight α	5.0e-4	30.2	24.8
	3.0e-4	26.7	15.0
	1.0e-4	**35.4**	**26.7**
PPL sampling interval	2.0-2.5	25.1	22.6
	2.5-3.0	**27.4**	**23.6**
	2.5-3.5	24.9	23.5
σ of Gaussian-guided PPL sampling	0.5	33.5	24.5
	0.25	**35.0**	**26.5**
	0.1	31.8	26.4

Table 3: Pass@1 scores of different configurations on AIME24 and AIME25 benchmarks.

the `ARN-1.1-SFT` model, despite the latter being trained on millions of trajectories.

4.4 Ablation Study

Q_1: Whether each setup of OXA has been empirically verified? A_1: Yes. Table 3 presents the results of ablating the hyperparameters used in OXA. By default, the experiments are conducted on the 1.5B model. The final configurations we use in the main experiment are bolded. Gaussian-guided PPL sampling outperforms the interval sampling, validating the benefit of mixing a small proportion of low-PPL data for better optimization.

Q_2: Can OXA models generalize to out-of-distribution reasoning tasks? A_2: Yes. We further evaluate OXA models on the PhD-level problems via GPQA diamond (Rein et al., 2023) and MMLU-STEM (Hendrycks et al., 2021). As illustrated in Figure 4 (d), OXA models consistently outperform SFT baselines, demonstrating that our method effectively enhances the model's fundamental complex reasoning capabilities.

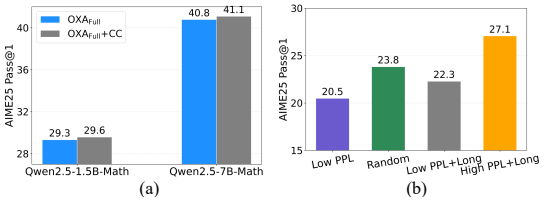

Figure 5: (a) Results of OXA_{Full} with Clip-Cov (OXA_{Full}+CC) on AIME25. (b) Comparison across different reasoning data selection strategies on AIME25. "Low PPL", "Random", and "High PPL + Long" correspond to SFT_{LP}, SFT, and OXA_{MLE}, respectively.

Q_3: Is OXA orthogonal to other methods? A_3: Yes. We combine OXA with Clip-Cov (Cui et al., 2025), which is an RLVR-enhanced method that controls entropy by restricting the update of high-covariance tokens to encourage exploration. Figure 5 (a) shows that OXA equipped with Clip-Cov achieves superior performance.

Q_4: Does OXA only benefit from selecting long data? A_4: No. Figure 5 (b) presents a comparison where the pre-trained LLM is fine-tuned on long data with low PPL. While this configuration yields marginal improvements over SFT_{LP}, it still lags significantly behind OXA_{MLE}, validating that the effectiveness of OXA stems from the integration of both low-confidence and long data.

5 Conclusion

We propose OXA to establish exploration-engaged initializations for the RLVR of mathematical LLMs. By leveraging MLE on low-confidence teacher data and unlikelihood training on high-confidence incorrect self-generated samples, OXA effectively boosts reasoning performance, expands the exploration space, and maintains high policy entropy. While validated on mathematics, OXA holds promise for other complex domains like code generation, which we defer to future work.

Limitations

Despite the performance gains, our work has two primary limitations. First, compared to vanilla SFT, OXA_{Full} incurs additional computational overhead due to the requirement of self-distillation reasoning trajectories. This sampling process is more resource-intensive than standard SFT. We further analyze the computation overhead of OXA in Appendix A.4. Second, due to limited computational resources, our empirical validation was restricted to models ranging from 1.5B to 7B parameters. However, we hypothesize that the benefits of OXA will be even more pronounced in larger-scale models. This is because larger models typically possess a greater capacity to internalize the reasoning paths with high PPL that smaller models might struggle to capture. We leave the exploration of OXA on larger model scales for future research.

References

Chenxin An, Zhihui Xie, Xiaonan Li, Ming Zhong, Shansan Gong, Lei Li, Jun Zhang, Jingjing Xu, and Lingpeng Kong. 2025. Long chain-of-thought fine-tuning via understanding-to-reasoning transition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34506–34522, Suzhou, China. Association for Computational Linguistics.

Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanovic, and Martin T. Vechev. 2025. Matharena: Evaluating llms on uncontaminated math competitions. *CoRR*, abs/2505.23281.

Hongye Cao, Zhixin Bai, Ziyue Peng, Boyan Wang, Tianpei Yang, Jing Huo, Yuyao Zhang, and Yang Gao. 2025. Efficient reinforcement learning with semantic and token entropy for llm reasoning. *arXiv preprint arXiv:2512.04359*.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with exploration: An entropy perspective. *CoRR*, abs/2506.14758.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *CoRR*, abs/2505.22617.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Benjamin Eysenbach and Sergey Levine. 2021. Maximum entropy RL (provably) solves some robust RL problems. *CoRR*, abs/2103.06257.

Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, and 31 others. 2025. Openthoughts: Data recipes for reasoning models. *CoRR*, abs/2506.04178.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Tianyi Jiang, Yi Bin, Yujuan Ding, Kainian Zhu, Fei Ma, Jingkuan Song, and Heng Tao Shen. 2025. Explore briefly, then decide: Mitigating LLM overthinking via cumulative entropy regulation. *CoRR*, abs/2510.02249.

Minwu Kim, Anubhav Shrestha, Safal Shrestha, Aadim Nepal, and Keith Ross. 2025. Reinforcement learning vs. distillation: Understanding accuracy and capability in LLM reasoning. *CoRR*, abs/2505.14216.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay V. Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Mengqi Liao, Xiangyu Xi, Ruinian Chen, Jia Leng, Yangen Hu, Ke Zeng, Shuai Liu, and Huaiyu Wan. 2025. Enhancing efficiency and exploration in reinforcement learning for llms. *CoRR*, abs/2505.18573.

Zihan Liu, Zhuolin Yang, Yang Chen, Chankyu Lee, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping.

2025. Acereason-nemotron 1.1: Advancing math and code reasoning through SFT and RL synergy. *CoRR*, abs/2506.13284.

Yijia Luo, Yulin Song, Xingyao Zhang, Jiaheng Liu, Weixun Wang, Gengru Chen, Wenbo Su, and Bo Zheng. 2025. Deconstructing long chain-of-thought: A structured reasoning optimization framework for long cot distillation. *CoRR*, abs/2503.16385.

Xingtai Lv, Yuxin Zuo, Youbang Sun, Hongyi Liu, Yuntian Wei, Zhekai Chen, Lixuan He, Xuekai Zhu, Kaiyan Zhang, Bingning Wang, Ning Ding, and Bowen Zhou. 2025. Towards a unified view of large language model post-training. *CoRR*, abs/2509.04419.

Yongyu Mu, Jiali Zeng, Bei Li, Xinyan Guan, Fandong Meng, Jie Zhou, Tong Xiao, and Jingbo Zhu. 2025. Dissecting long reasoning models: An empirical study. *CoRR*, abs/2506.04913.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems, EuroSys 2025, Rotterdam, The Netherlands, 30 March 2025 - 3 April 2025*, pages 1279–1297. ACM.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, and 75 others. 2025. Kimi k1.5: Scaling reinforcement learning with llms. *CoRR*, abs/2501.12599.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, Haosheng Zou, Yongchao Deng, Shousheng Jia, and Xiangzheng Zhang. 2025. Light-r1: Curriculum sft, DPO and RL for long COT from scratch and beyond. *CoRR*, abs/2503.10460.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8:229–256.

Ronald J Williams and Jing Peng. 1991. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268.

Bingquan Xia, Bowen Shen, Cici, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, Liang Zhao, Peidian Li, Peng Wang, Shihua Yu, Shimao Chen, Weikun Wang, Wenhan Ma, Xiangwei Deng, Yi Huang, and 44 others. 2025. Mimo: Unlocking the reasoning potential of language model - from pretraining to posttraining. *CoRR*, abs/2505.07608.

Hongling Xu, Qi Zhu, Heyuan Deng, Jinpeng Li, Lu Hou, Yasheng Wang, Lifeng Shang, Ruifeng Xu, and Fei Mi. 2025. KDRL: post-training reasoning llms via unified knowledge distillation and reinforcement learning. *CoRR*, abs/2506.02208.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. 2025. Learning to reason under off-policy guidance. *CoRR*, abs/2504.14945.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025a. Qwen3 technical report. *CoRR*, abs/2505.09388.

Cehao Yang, Xueyuan Lin, Chengjin Xu, Xuhui Jiang, Xiaojun Wu, Honghao Liu, Hui Xiong, and Jian Guo. 2025b. Select2reason: Efficient instruction-tuning data selection for long-cot reasoning. *CoRR*, abs/2505.17266.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *CoRR*, abs/2504.13837.

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

Xiaoyun Zhang, Xiaojian Yuan, Di Huang, Wang You, Chen Hu, Jingqing Ruan, Kejiang Chen, and Xing Hu. 2025b. Rediscovering entropy regularization: Adaptive coefficient unlocks its potential for LLM reinforcement learning. *CoRR*, abs/2510.10959.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark W. Barrett, and Ying Sheng. 2024a. Sglang: Efficient execution of structured language model programs. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024b. Llamafactory: Unified efficient fine-tuning of 100+ language models. *CoRR*, abs/2403.13372.

A Appendix

A.1 Theoretical Analysis of Unlikelihood Loss

In this section, we analyze the gradient dynamics of the Unlikelihood Loss ($\mathcal{L}_{\text{UL}}$) compared to the standard Cross-Entropy Loss ($\mathcal{L}_{\text{CE}}$) to justify the necessity of a small scaling factor α.

Let $z \in \mathbb{R}^V$ denote the logit vector output by the model at a specific time step t, where V is the vocabulary size. Let $p_k = \text{softmax}(z)_k$ represent the predicted probability for token k. For a given input context $s_{<t}$, let x_t be the specific token index targeted by the loss function (the ground truth token for $\mathcal{L}_{\text{CE}}$ or the negative token to be penalized for $\mathcal{L}_{\text{UL}}$).

A.1.1 Gradient of Cross-Entropy Loss

The gradient of $\mathcal{L}_{\text{CE}}$ with respect to any logit z_j is bounded. Specifically:

$$\frac{\partial \mathcal{L}_{\text{CE}}}{\partial z_j} = p_j - \mathbb{1}(j = x_t), \tag{6}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Since $p_j \in (0, 1)$, the gradient magnitude is strictly bounded such that $\left|\frac{\partial \mathcal{L}_{\text{CE}}}{\partial z_j}\right| < 1$. This ensures stable updates regardless of the model's current confidence.

A.1.2 Gradient of Unlikelihood Loss

The unlikelihood objective aims to minimize the probability of a negative token x_t. The loss is defined as $\mathcal{L}_{\text{UL}} = -\log(1 - p_{x_t})$. Using the chain rule, we derive the gradient with respect to the logits.

For the target negative token logit (z_{x_t}):

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{UL}}}{\partial z_{x_t}} &= \frac{\partial \mathcal{L}_{\text{UL}}}{\partial p_{x_t}} \cdot \frac{\partial p_{x_t}}{\partial z_{x_t}} \\ &= \frac{1}{1 - p_{x_t}} \cdot p_{x_t}(1 - p_{x_t}) = p_{x_t}. \end{aligned} \tag{7}$$

This term is bounded within $[0, 1]$. However, the instability arises from the gradients with respect to *other* tokens z_j (where $j \neq x_t$). The derivative of the softmax function for off-target indices is $\frac{\partial p_{x_t}}{\partial z_j} = -p_{x_t} p_j$. Thus:

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{UL}}}{\partial z_j} &= \frac{\partial \mathcal{L}_{\text{UL}}}{\partial p_{x_t}} \cdot \frac{\partial p_{x_t}}{\partial z_j} \\ &= \frac{1}{1 - p_{x_t}} \cdot (-p_{x_t} p_j) \\ &= -p_j \cdot \underbrace{\left(\frac{p_{x_t}}{1 - p_{x_t}}\right)}_{\text{Odds Ratio Term}}. \end{aligned} \tag{8}$$

A.1.3 Instability and Weight Scaling

Equation 8 reveals a critical instability mechanism. The gradient for all non-target logits is scaled by the odds ratio of the negative token, $\frac{p_{x_t}}{1-p_{x_t}}$.

High-Confidence Errors. Consider a scenario where the model assigns a high probability to a hallucinated or incorrect token x_t (e.g., $p_{x_t} = 0.99$). In this case, the gradient scaling factor becomes:

$$\frac{0.99}{1 - 0.99} = 99. \tag{9}$$

Consequently, the gradient applied to all other logits z_j is amplified by a factor of roughly 100 compared to standard training dynamics. As $p_{x_t} \to 1$, this term approaches infinity.

If the scaling factor α in the final objective $\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha \cdot \mathcal{L}_{\text{UL}}$ is set to 1.0, these exploded gradients from high-confidence errors effectively overwrite the semantic knowledge learned via $\mathcal{L}_{\text{CE}}$, leading to catastrophic forgetting or model divergence. By setting α to a small value (e.g., 10^{-4}), we counteract the explosion of the odds ratio term, ensuring that the unlikelihood updates remain comparable in magnitude to the maximum likelihood updates, thus stabilizing the training process.

A.2 Preliminary Experiments

As illustrated in Figure 6, the performance of models fine-tuned on data subsets from different PPL ranges exhibits significant variance. Specifically,

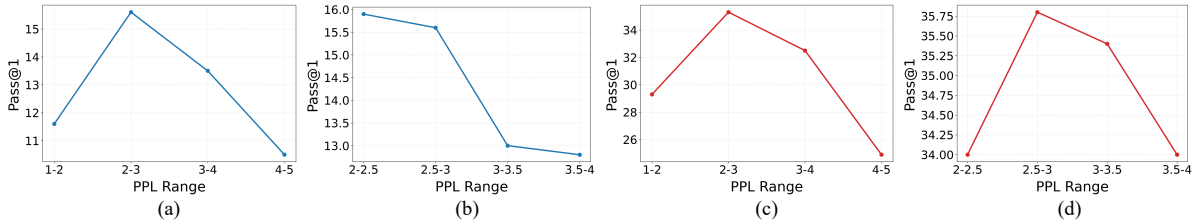

Figure 6: Impact of data perplexity ranges on reasoning performance. (a)-(b) Performance on AIME24 of 1.5B models fine-tuned on various PPL intervals. (c)-(d) Corresponding results for 7B models across different PPL ranges.

training on data with excessively low PPL (e.g., 1.0–2.0) yields suboptimal reasoning capabilities, as these samples likely represent patterns the model has already mastered, offering limited signal for further improvement. In contrast, data with slightly higher PPL (e.g., 2.0–3.0) achieves the best performance, as it trains the model to internalize previously uncaptured reasoning trajectories.

However, we also observe that if the PPL is too high, the complexity of the reasoning paths may exceed the model's current capacity, leading to optimization difficulties. To strike an optimal balance between learning and exploration potential, we propose Gaussian-guided PPL sampling. This approach allows us to concentrate training on the most effective PPL regions while maintaining a smooth distribution. Based on these preliminary findings, we set the Gaussian mean μ to 2.5 for 1.5B models and 3.0 for 7B models in our main experiments.

A.3 Detailed Experimental Setup

Model configurations. To enable long-sequence modeling for mathematical reasoning, we adjust the rotary positional embedding (RoPE) parameters for the Qwen2.5-1.5B-Math and Qwen2.5-7B-Math models. Specifically, we increase rope theta from $10,000$ to $1,000,000$ and extend max position embeddings from $4,096$ to $40,000$. For `LLaMA3.2-3B` and `Qwen3-1.7B`, no modifications are required as their native context window of $32,768$ is sufficient for our experiments. Additionally, we remove the system prompt component from the tokenizer templates across all models to ensure a consistent and simplified input format.

Data preparation. The first objective of OXA involves sampling from teacher-distilled SFT data. For `Qwen2.5-1.5B-Math`, the sampling hyperparameters are set as follows: `MIN PPL` = 1.0, `MAX PPL` = 5.0, `BIN WIDTH` = 0.05, `TARGET STD` = 0.25, and `TOTAL SAMPLES` = $50,000$. We maintain these settings for other models while adjusting the `TARGET CENTER`: it is set to 2.5 for `Qwen2.5-1.5B-Math` and `LLaMA3.2-3B`, and 3.0 for `Qwen2.5-7B-Math` and `Qwen3-1.7B`. To elicit structured chain-of-thought reasoning, we append the following instruction to each SFT problem: "\nLet's reason step by step. Enclose the reasoning process within <think>...</think>, then summarize it and present the final answer within \boxed — for example: <think>reasoning process here</think> \boxedanswer here." For RLVR training, we use `DeepSeek-Distill-Qwen2.5-7B` to perform 8-sample generation per query. We select $10,000$ trajectories with pass rates between 0.2 and 0.8, effectively filtering out tasks that are either trivial or excessively difficult.

Training details. For SFT, the Qwen2.5-1.5B-Math model is trained with a cutoff len of $32,768$, a learning rate of 2.5×10^{-4}, and 6.0 epochs. We use a warmup ratio of 0.03, weight decay of 0.1, and Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.95$, using a global batch size of 128. For other models, we adjust only the learning rates: 5.0×10^{-5} for Qwen2.5-7B-Math, and 2.0×10^{-4} for both `LLaMA3.2-3B` and `Qwen3-1.7B`. Reinforcement learning parameters are kept uniform: train batch size is 64, max response length is $16,384$, actor learning rate is 2.0×10^{-6}, KL coefficient is 0.001, with 8 rollouts per query at a temperature of 0.85. In our data scaling experiments with `Qwen2.5-7B-Math`, we increase the global batch size to 384 and the learning rate to 1.5×10^{-4} to ensure the total number of optimization updates remains consistent with our baseline experiments. We use `LLaMA-Factory` (Zheng et al., 2024b) and `Verl` (Sheng et al., 2025) for fine-tuning and reinforcement learning, respectively.

A.4 Compute Overhead Discussion

Compared to conventional SFT, the additional computational overhead of OXA primarily stems from two components: PPL estimation for data selection and model decoding during self-distillation.

Efficiency of PPL estimation. While OXA introduces a PPL calculation step, this process is highly efficient in practice. Since PPL estimation is performed through a single forward pass, the negative log-likelihoods of all tokens in a sequence are computed simultaneously in parallel. This allows the selection process to scale efficiently with sequence length, avoiding the sequential bottlenecks typical of autoregressive generation.

Self-distillation and mitigation. The primary source of extra compute relative to vanilla SFT is the generation of high-confidence error trajectories via self-distillation. However, the cost of this phase is significantly mitigated by modern inference-time optimizations. In our pipeline, we leverage high-throughput inference frameworks such as vLLM (Kwon et al., 2023) and SGLang (Zheng et al., 2024a), coupled with model quantization techniques. These advancements ensure that the generation of self-distilled data is both rapid and cost-effective, making OXA a practical choice for large-scale training.

A.5 Comparison OXA Fine-Tuning with Direct Preference Optimization

While both OXA and direct preference optimization (DPO) (Rafailov et al., 2023) promote desirable samples and suppress undesirable ones, they differ fundamentally in two key aspects:

Data structure and coupling. DPO is inherently a pairwise framework, requiring each training instance to consist of a triplet: a single query associated with both a "chosen" (desirable) and a "rejected" (undesirable) response. This constraint limits the utilization of unpaired data. In contrast, the data requirements for OXA are entirely decoupled. OXA permits queries to be associated with only a desirable or only an undesirable response, significantly lowering the barrier for data collection. Furthermore, this decoupling allows for flexible control over the mixing ratio of desirable and undesirable samples within each training batch, a hyperparameter that can be tuned to balance exploration and exploitation.

Optimization mechanism. The training dynamics of the two methods are distinct. DPO employs a contrastive loss, which primarily focuses on maximizing the relative log-probability gap between the chosen and rejected responses. Conversely, OXA treats promotion and suppression as two independent objectives: the NLL loss focuses on internalizing correct reasoning patterns, while the unlikelihood loss directly minimizes the probability of incorrect paths. There is no intrinsic mathematical linkage between these two objectives in OXA, allowing the model to learn from each type of signal independently without the need for a direct comparison between two specific trajectories for every query.