

Beyond Rejection Sampling: Trajectory Fusion for Scaling Mathematical Reasoning

Jie Deng¹, Hanshuang Tong¹, Jun Li^{1*}, Shining Liang¹,
Ning Wu¹, Hongzhi Li¹, Yutao Xie¹

¹Microsoft, Beijing, China

dengjie530@outlook.com, junli1@microsoft.com

Abstract

Large language models (LLMs) have made impressive strides in mathematical reasoning, often fine-tuned using rejection sampling, which retains only correct reasoning trajectories. While effective, this paradigm treats supervision as a binary filter that systematically excludes teacher-generated errors, leaving a gap in how reasoning failures are modeled during training. In this paper, we propose **TrajFusion**, a fine-tuning strategy that reframes rejection sampling as a structured supervision construction process. Specifically, TrajFusion forms fused trajectories that explicitly model trial-and-error reasoning by interleaving selected incorrect trajectories with reflection prompts and correct trajectories. The length of the fused sample is adaptively controlled based on the frequency and diversity of teacher errors, providing richer supervision for challenging problems while safely reducing to vanilla rejection sampling fine-tuning (RFT) when error signals are uninformative. TrajFusion requires no changes to the architecture or training objective. Extensive experiments across multiple math benchmarks demonstrate that TrajFusion consistently outperforms RFT, particularly on challenging and long-form reasoning problems.

1 Introduction

Large language models (LLMs) have made significant progress in mathematical reasoning, facilitated by techniques such as chain-of-thought prompting (Wei et al., 2022) and supervised fine-tuning on high-quality solution trajectories (Cobbe et al., 2021; Yu et al., 2023). When combined with strong teacher models, these approaches enable LLMs to acquire increasingly sophisticated problem-solving behaviors and deliver exceptional accuracy across diverse mathematical benchmarks.

A widely adopted method for constructing supervised reasoning data is *rejection sampling* (Yuan

et al., 2023b). In this approach, a teacher model generates multiple candidate trajectories for a given problem, retaining only those with correct final answers. Rejection sampling fine-tuning (RFT) then performs supervised training on this filtered set of correct trajectories. RFT is appealing due to its simplicity, requiring no human annotation, and its demonstrated ability to improve model performance across a wide range of mathematical tasks (Yuan et al., 2023a; Tong et al., 2024; Xiong et al., 2025). Conceptually, RFT treats supervision as a binary filtering process, where correctness alone determines whether a trajectory contributes to learning.

However, this paradigm has a fundamental limitation: it systematically discards all teacher-generated incorrect trajectories, despite the fact that such solutions contain valuable diagnostic information about reasoning failures. Incorrect trajectories reveal common computational slips, missing assumptions, misapplied formulas, and logically plausible yet invalid argument paths. These failure patterns are not noise; rather, they provide informative supervision that can help models recognize and avoid common reasoning errors, especially on long or ambiguous problems where distinguishing valid from flawed reasoning is essential (An et al., 2023; Pan et al., 2025). This leads to our central insight: ***teacher-generated errors should not be discarded, but incorporated as structured supervision.***

Building on this insight, we propose **TrajFusion**, a simple yet effective fine-tuning framework that extends RFT by incorporating selected incorrect reasoning trajectories into the supervision signal. For each problem, TrajFusion constructs a single training sample by interleaving a selected subset of incorrect trajectories with the correct one, linked by brief reflection prompts. This approach explicitly simulates a trial-and-error reasoning process, exposing the model to both typical failure modes and their resolutions. The selection in TrajFusion

*Corresponding author.

is fully adaptive. When a problem yields diverse or informative errors, more incorrect trajectories are incorporated to enrich supervision; when error signals are sparse or uninformative, TrajFusion provably degenerates to vanilla RFT. In our method, error diversity is treated as an explicit control signal rather than an implicit assumption, which allows the method to behave robustly across problems of varying difficulty. Importantly, TrajFusion requires no additional sampling budget, no auxiliary models or critique generators, no architectural modifications, and no external difficulty labels.

We evaluate TrajFusion on two base models, LLaMA3-8B and DeepSeekMath-7B, across six mathematical reasoning benchmarks covering both in-domain and out-of-domain settings. Under both low-data (15K) and full-data (100K) regimes, TrajFusion consistently outperforms RFT, with particularly large gains on challenging datasets such as MATH, TheoremQA, and OlympiadBench. To assess its scalability, we further apply TrajFusion to long-form reasoning data distilled from DeepSeek-R1-Distill-Qwen-7B and fine-tune a 32K-context version of Qwen2.5-Math-7B. TrajFusion again surpasses RFT, achieving higher accuracy on competition-level benchmarks including AIME24, AIME25, and MATH-500.

Contributions.

- We formalize a previously underexplored supervision gap in RFT: the systematic exclusion of teacher-generated incorrect trajectories despite their diagnostic value for modeling reasoning failures. This motivates treating such errors as valuable supervision, rather than noise.
- We propose **TrajFusion**, a structured supervision framework that reframes RFT from a binary filtering mechanism into a supervision construction process, adaptively fusing incorrect and correct reasoning trajectories through lightweight reflections. An adaptive mechanism automatically adjusts the amount of fused content based on the diversity and informativeness of teacher errors, while preserving standard RFT behavior when errors provide limited value.
- We demonstrate that TrajFusion consistently and scalably improves over RFT across multiple models, datasets, data regimes, and context lengths, with notable gains on challenging

benchmarks and long-context mathematical reasoning.

2 Preliminaries

2.1 Rejection Sampling Fine-Tuning

Rejection sampling fine-tuning (RFT) constructs supervised data by sampling multiple candidate trajectories for a problem x ,

$$\mathcal{Y}(x) = \{y_1, \dots, y_K\}, \quad y_k \sim T(\cdot | x),$$

and retaining only those with correct final answers,

$$\mathcal{Y}^+(x) = \{y \in \mathcal{Y}(x) \mid \text{corr}(y) = 1\}.$$

All incorrect trajectories,

$$\mathcal{Y}^-(x) = \mathcal{Y}(x) \setminus \mathcal{Y}^+(x),$$

are discarded. While this produces a clean training signal, it also removes the full distribution of reasoning attempts. We analyze these sampling distributions to understand what information RFT leaves unused.

2.2 Sampling Distributions

We analyze the sampling behavior of Qwen2.5-Math-7B-Instruct (Yang et al., 2024) on the DeepMath dataset (He et al., 2025) by examining error rates and the diversity of incorrect answers across sampled trajectories. Our analysis focuses on problems with mixed outcomes, where both correct and incorrect trajectories are observed.

Figure 1 summarizes the sampling statistics. The left panel shows the distribution of error rates across problems. Approximately 40% of problems have near-zero error rates and are almost always solved correctly, while about 18% are nearly always answered incorrectly. Between these extremes lies a substantial set of problems with mixed outcomes, where the model alternates between success and failure across attempts; these intermediate cases are particularly informative for analyzing reasoning behavior under uncertainty.

The right panel characterizes failure structure by measuring the diversity of incorrect final answers using Shannon entropy. Error diversity increases with error rate ($\rho = 0.55$, $p < 10^{-10}$), indicating that harder problems tend to induce a broader range of distinct failure modes rather than repeated identical mistakes.

Beyond this overall trend, two characteristic patterns emerge. Some problems exhibit **low error**

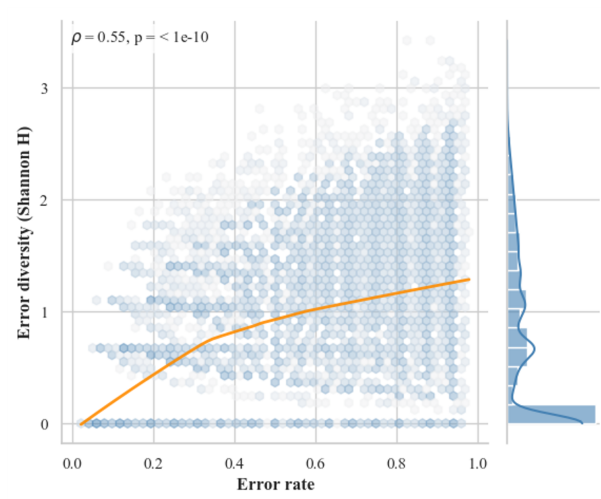
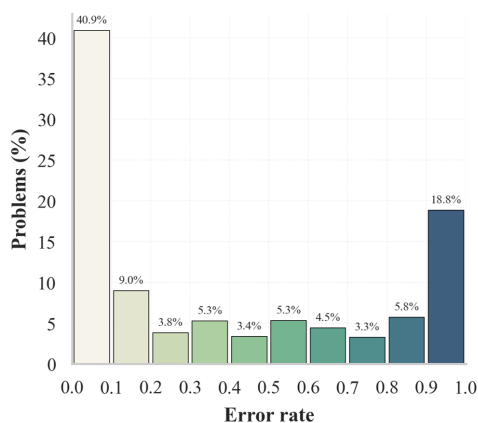


Figure 1: Sampling statistics for Qwen2.5-Math-7B-Instruct on the DeepMath dataset. Left: distribution of error rates across problems. Right: diversity of incorrect final answers measured by Shannon entropy

rates but high diversity among incorrect answers. In these cases, the model usually follows a reliable reasoning path, while rare failures arise from unstable or divergent intermediate steps. Other problems show **high error rates but low diversity**, where the model fails frequently yet converges to nearly the same incorrect answer each time. This behavior indicates a systematic misunderstanding or a strong inductive bias that repeatedly drives the model toward a specific incorrect solution.

Taken together, these observations demonstrate that the discarded set $\mathcal{Y}^-(x)$ encodes rich structural information beyond a binary notion of correctness. Incorrect trajectories vary not only in frequency but also in qualitative structure, reflecting distinct modes of reasoning instability and persistent misconceptions.

3 Method

Building on the observations in Section 2, we introduce **TrajFusion**, a structured fine-tuning approach that integrates selected incorrect reasoning trajectories together with a correct one into a single fused training sample. Instead of relying exclusively on the correct solution, TrajFusion leverages information from diverse failure modes to provide richer supervision, while remaining fully compatible with vanilla RFT pipelines.

3.1 Sampling and Scoring

For each problem x , we sample K reasoning trajectories

$$\mathcal{Y}(x) = \{y_1, \dots, y_K\}, \quad y_k \sim T(\cdot | x),$$

and partition them into correct and incorrect subsets using an automatic verifier, yielding $\mathcal{Y}^+(x)$ and $\mathcal{Y}^-(x)$ respectively.

To characterize the sampling outcomes at the problem level, we compute two statistics. The first is an error-rate estimate

$$r(x) = \frac{|\mathcal{Y}^-(x)|}{K},$$

which measures how frequently the teacher model produces incorrect trajectories among the sample attempts for problem x .

The second statistic quantifies the diversity of incorrect outcomes:

$$u(x) = |\{\text{Ans}(y) : y \in \mathcal{Y}^-(x)\}|.$$

Here distinct final answers represent different error types, allowing us to distinguish between problems where the model collapses to a single failure mode and those where it explores multiple incorrect reasoning paths.

3.2 Difficulty- and Diversity-Aware Selection

TrajFusion selectively incorporates incorrect trajectories into training. The selection is jointly guided by error frequency and diversity: incorrect trajectories are included only when they occur with sufficient frequency, and a larger number of incorrect trajectories are selected as failures become more diverse.

The number of retained incorrect trajectories is determined by

$$k(x) = \min(k_{\max}, \lfloor \alpha r(x) u(x) \rfloor),$$

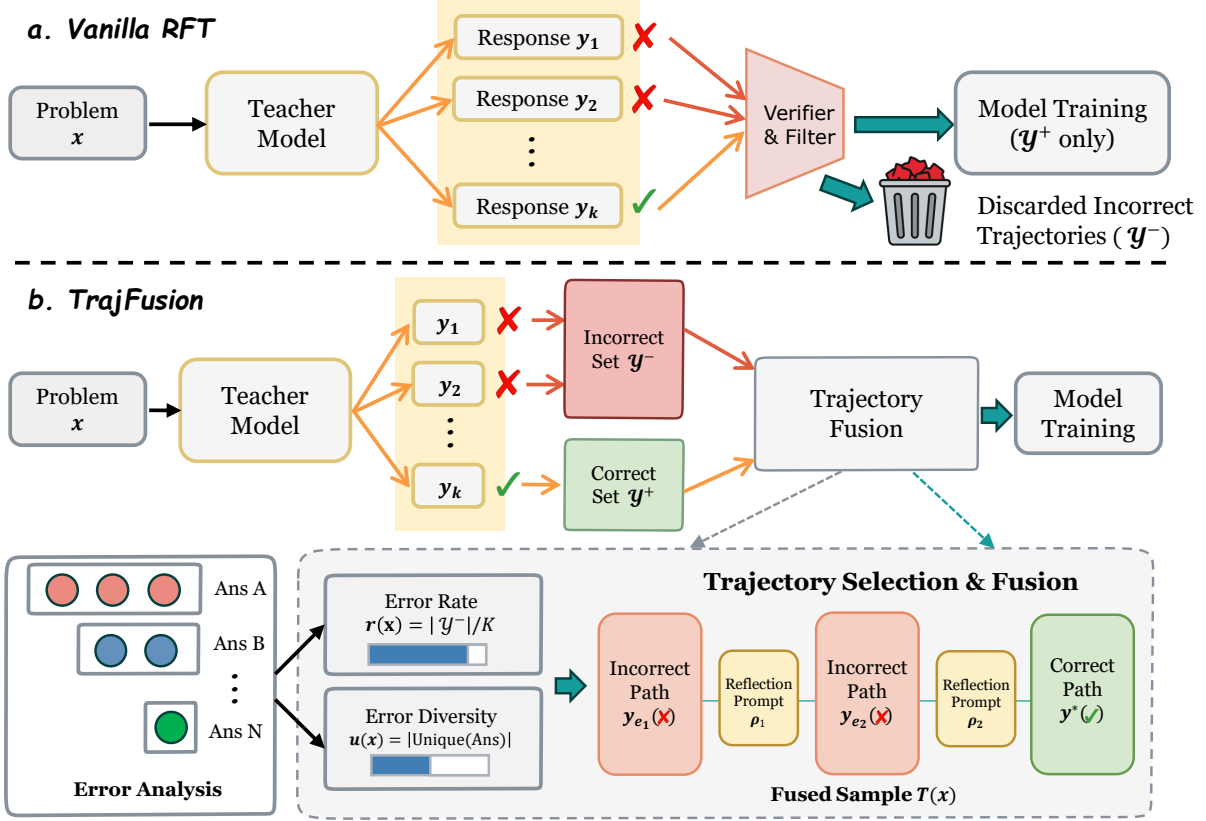


Figure 2: Comparison between vanilla rejection-sampling fine-tuning (RFT) and our TrajFusion framework. **Top:** Vanilla RFT samples multiple responses from a teacher model and retains only verified correct trajectories (\mathcal{Y}^+) for training, discarding all incorrect ones. **Bottom:** TrajFusion explicitly separates correct (\mathcal{Y}^+) and incorrect (\mathcal{Y}^-) trajectories, performs problem-level error analysis, and constructs fused training trajectories by selectively integrating informative incorrect reasoning paths with corrected trajectories.

where k_{\max} is a fixed upper bound and α controls the sensitivity to error diversity. This rule yields an adaptive behavior:

- When no incorrect trajectory is observed, $r(x) = 0$ and thus $k(x) = 0$, reducing TrajFusion to vanilla RFT without introducing any additional error supervision.
- When errors are present, $k(x)$ increases with both the error rate $r(x)$ and the answer diversity $u(x)$: homogeneous failures (small $u(x)$) result in fewer selected incorrect trajectories, while higher error rates and more diverse outcomes lead to richer supervision.

To obtain representative yet compact supervision, we organize incorrect trajectories $\mathcal{Y}^-(x)$ by their final answers and prioritize error groups by frequency. Within each selected group, we favor shorter trajectories as representatives, encouraging concise supervision while preserving diversity across failure outcomes.

3.3 Fused Sample Construction

Given a set of selected incorrect trajectories $\{y_{e_1}, \dots, y_{e_k}\}$ and a correct trajectory $y^* \in \mathcal{Y}^+(x)$, TrajFusion constructs a single fused training sample

$$T(x) = [y_{e_1}, \rho_1, \dots, y_{e_k}, \rho_k, y^*],$$

where each ρ_i is a fixed reflection prompt that explicitly marks the transition between successive attempts and provides contextual signals for correction.

The training of TrajFusion proceeds with standard supervised fine-tuning on the fused sequence sample. The model is optimized via next-token prediction using the negative log-likelihood objective

$$\mathcal{L}_{SFT} = - \sum_{t=1}^{|T(x)|} \log p_{\theta}(T_t | T_{<t}),$$

without any modification to the model architecture or training objective. TrajFusion therefore operates as a supervision construction mechanism, that

enriches the learning signal while maintaining compatibility with existing fine-tuning pipelines.

4 Experiments

4.1 Implementation Details

Dataset Construction and Sampling. We construct two training sets: an initial 15K set and an expanded 100K set. The 15K set is drawn from GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), and AIME (pre-2024). To introduce additional variety, we incorporate more problems from the DeepMath dataset (He et al., 2025), forming the 100K set.

For each problem x , we sample $K = 16$ chain-of-thought trajectories using Qwen2.5-Math-7B-Instruct (Yang et al., 2024) with temperature 0.7 and top- p 0.95, generated through vLLM (Kwon et al., 2023) with a 4096-token limit. An automatic verifier checks final answers, and queries without any correct solution are excluded from training.

Training Setup. We fine-tune two base models, LLaMA3-8B-base (Dubey et al., 2024) and DeepSeek-Math-7B-base (Shao et al., 2024) using ms-swift framework (Zhao et al., 2024). Training uses a batch size of 64, learning rate 5×10^{-5} , and sequence packing with ZeRO-1 optimization (Rasley et al., 2020). The 15K set is trained for 5 epochs and the 100K set for 3 epochs.

Evaluation Benchmarks. We evaluate on a suite of mathematical reasoning benchmarks, including GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), CollegeMath (Tang et al., 2024), DeepMind-Mathematics (Saxton et al., 2019), OlympiadBench-Math (He et al., 2024), and TheoremQA (Chen et al., 2023). All scores report Pass@1 accuracy.

Baselines. We compare TrajFusion with representative data construction and fine-tuning strategies for mathematical reasoning, including MMIQC (Liu et al., 2025), DART-Math (Tong et al., 2024), MathFusion (Pei et al., 2025), DeepSeekMath-7B-Instruct (Shao et al., 2024), and LEMMA (Pan et al., 2025). As our primary controlled baseline, we include *vanilla RFT*, which fine-tunes only on the correct trajectories obtained from rejection sampling. This allows us to isolate the contribution of incorporating selected incorrect reasoning paths.

4.2 Main Results

Overall performance and data efficiency. Table 1 summarizes the performance of TrajFusion across six mathematical reasoning benchmarks. With only 15k training examples, TrajFusion already achieves competitive results on both backbones. On DeepSeekMath-7B, it reaches 55.8% on MATH and 48.8% on average, outperforming several data-construction baselines that use substantially larger training sets (e.g., MathFusion at 60k and MMIQC at 2.3M). This demonstrates that TrajFusion can extract meaningful supervisory signals even in low-data regimes.

At 100k examples, TrajFusion achieves the highest average accuracy among all compared training strategies: 41.3% on LLaMA3-8B and 52.3% on DeepSeekMath-7B. The method continues to yield improvements as data scales, indicating that trajectory-level supervision from both correct and selected incorrect attempts remains beneficial in larger training regimes.

Effectiveness varies with reasoning complexity. TrajFusion provides the largest gains on benchmarks characterized by longer or more structured reasoning, such as MATH, OlympiadBench, and TheoremQA. These tasks naturally produce more diverse incorrect attempts, aligning with the model’s strength in leveraging trajectory-level variation. At 100k examples, TrajFusion improves over DARTMath by +4.8 points on MATH for LLaMA3-8B and +5.5 points for DeepSeekMath-7B, and achieves the highest accuracy on TheoremQA across all methods.

On GSM8K, TrajFusion performs similarly to other approaches, with smaller gains than on the more complex benchmarks. GSM8K problems involve short, well-structured reasoning chains with limited variability, resulting in fewer opportunities for trajectory fusion to provide additional benefit.

Comparison with vanilla RFT. To isolate the contribution of integrating incorrect trajectories, the vanilla RFT baseline uses the same sampled queries and the same correct trajectories as TrajFusion, but discards all incorrect attempts. Across both model backbones and training sizes, TrajFusion consistently outperforms vanilla RFT. For LLaMA3-8B, the improvements are +6.1 points at 15k and approximately +3 points at 100k. For DeepSeekMath-7B, the gains are +3.0 and +2.1 points respectively.

Table 1: **Pass@1 accuracy on six mathematical reasoning benchmarks.** Bold numbers indicate the best performance under the same base model. TrajFusion consistently outperforms vanilla RFT and prior data construction methods across both model families and data scales.

Method	Train Size	Mathematical Reasoning Benchmarks						AVG
		MATH	GSM8K	College	DM	Olympiad	Theorem	
<i>LLaMA3-8B Base Model</i>								
ICL	—	21.2	51.0	19.9	27.4	4.2	19.8	23.9
MetaMath	400k	32.5	77.3	20.6	35.0	5.5	13.8	30.8
MMIQc	2.3M	39.5	77.6	29.5	41.0	9.6	16.2	35.6
DARTMath	590k	46.6	81.1	28.8	48.0	14.5	19.4	<u>39.7</u>
MathFusion	60k	46.5	<u>79.2</u>	27.9	43.4	<u>17.2</u>	20.0	39.0
LEMMA	89k	38.3	<u>79.2</u>	24.7	—	—	—	—
Vanilla RFT	15k	29.5	72.3	15.8	28.4	6.4	14.9	27.9
	100k	<u>48.5</u>	76.9	<u>28.9</u>	41.1	14.5	<u>22.6</u>	38.8
TrajFusion	15k	36.6	<u>79.2</u>	21.6	37.9	8.6	20.0	34.0
	100k	51.4	77.3	31.6	<u>43.7</u>	20.2	23.3	41.3
<i>DeepSeekMath-7B Base Model</i>								
ICL	—	35.5	64.2	34.7	45.2	9.3	23.5	35.4
Instruct	780k	46.9	<u>82.7</u>	37.1	52.2	14.2	28.1	43.5
MMIQc	2.3M	45.3	79.0	35.3	52.9	13.0	23.4	41.5
DARTMath	590k	53.6	86.8	<u>40.7</u>	61.6	21.7	32.2	49.4
MathFusion	60k	53.4	77.9	39.8	65.8	23.3	24.6	47.5
LEMMA	89k	50.6	80.4	35.6	—	—	—	—
Vanilla RFT	15k	51.8	78.5	38.3	64.9	19.6	21.4	45.8
	100k	<u>56.2</u>	79.2	40.1	<u>69.1</u>	<u>25.6</u>	31.2	<u>50.2</u>
TrajFusion	15k	55.8	81.4	39.0	66.8	23.1	26.9	48.8
	100k	59.1	82.5	41.2	69.6	27.0	34.1	52.3

Gains are especially pronounced on structurally challenging benchmarks. On OlympiadBench, TrajFusion improves LLaMA3-8B from 6.4% to 8.6%, and DeepSeekMath-7B from 19.6% to 23.1% at 15k. On TheoremQA, improvements of 5–6 points are observed on both backbones. These results support the central hypothesis of TrajFusion: selectively incorporating informative incorrect reasoning trajectories provides richer supervision than correctness-only training, yielding consistent benefits in both low- and high-data settings.

4.3 Effect of Error Supervision

We study how the quantity and diversity of fused incorrect trajectories affect model performance. Table 2 reports accuracy, the average number of fused incorrect paths per training example (**#Err**), and average generated output length (**Tokens**) on GSM8K and MATH using LLaMA3-8B.

Effect of error quantity. Varying k_{\max} controls the maximum number of incorrect trajectories that may be included in a fused sequence. As shown

Table 2: Ablation on the effect of fusion count (k_{\max}) and error diversity on **GSM8K** and **MATH** using **LLaMA3-8B**. **#Err** denotes the average number of fused incorrect trajectories per example.

Strategy	#Err	GSM8K		MATH	
		Acc	Tokens	Acc	Tokens
TF($k_{\max}=1$)	0.62	72.0	125	39.6	912
TF($k_{\max}=2$)	1.02	72.9	145	41.9	1056
TF($k_{\max}=4$)	1.49	74.8	155	44.2	1173
TF($k_{\max}=6$)	1.74	74.0	153	44.1	1192
TF(<i>LowDiv</i>)	1.13	72.3	291	40.7	1603
TF(<i>HighDiv</i>)	1.13	74.0	120	43.2	1056

in the top block of Table 2, increasing k_{\max} from 1 to 4 raises both **#Err** and performance on both benchmarks. For instance, MATH accuracy improves from 39.56% to 44.18%, and GSM8K from 71.95% to 74.83%. These results suggest that a moderate amount of error supervision provides additional informative signals beyond the correct trajectory alone. However, increasing k_{\max} further to

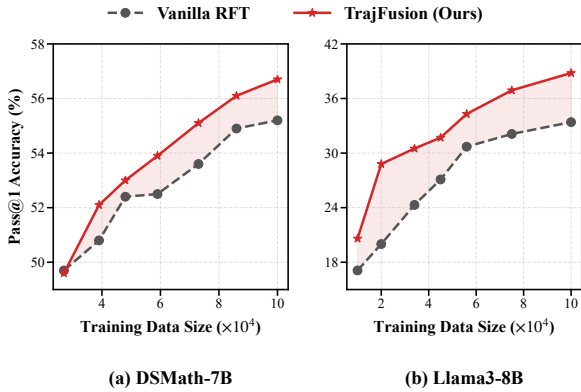


Figure 3: Pass@1 accuracy on MATH during the first training epoch, evaluated at fixed steps. TrajFusion (red) consistently outperforms Vanilla RFT (gray) across both models. The shaded areas highlight the performance margin, with a significant gap established early in training.

6 yields no additional benefit despite producing more fused errors, indicating diminishing returns when too many incorrect paths are included.

Effect of error diversity. We next control for the number of fused errors ($\#Err=1.138$) and vary only the selection strategy. **LowDiv** samples errors from the most frequent incorrect answer bucket, whereas **HighDiv** prefers selecting trajectories from different incorrect outcomes. As shown in the bottom block of Table 2, HighDiv consistently outperforms LowDiv (+1.67% on GSM8K and +2.50% on MATH). HighDiv also produces substantially shorter outputs. This indicates that redundant error patterns are less beneficial, while structurally distinct incorrect trajectories provide more useful supervision and lead to more concise reasoning behavior at inference time.

4.4 Analysis

Data Efficiency. Figure 3 shows the performance dynamics throughout the first epoch, evaluated at fixed intervals to track how accuracy changes with increasing data exposure. In both DeepSeek-Math-7B and LLaMA3-8B model, TrajFusion consistently surpasses the Vanilla RFT baseline across all data intervals. For LLaMA3-8B specifically, a distinct performance gap emerges as early as 20k samples and persists throughout the training process. The shaded regions in the plots illustrate this robust margin, indicating that TrajFusion achieves higher accuracy than the baseline given the same number of training samples. These observations suggest that the additional supervision from incor-

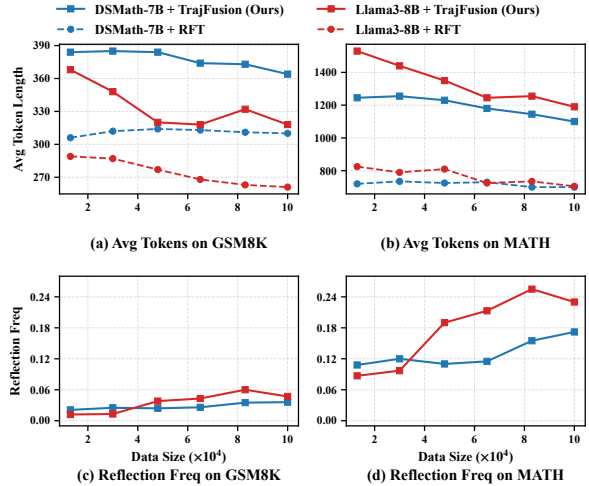


Figure 4: Average output length and reflection frequency on GSM8K and MATH. (a, b) TrajFusion initially results in longer generations, but the average length gradually decreases as training progresses. (c, d) The frequency of reflection tokens shows an upward trend, particularly on the more complex MATH dataset.

rect paths and reflections contributes to better sample efficiency.

Token Efficiency and Reflection. We track the average output token length and reflection frequency to examine the generation behavior of TrajFusion (Figure 4). Initially, TrajFusion models produce longer sequences than the baselines due to the inclusion of trial-and-error paths. However, we observe a continuous decrease in average token count throughout the training process, suggesting that the generated trajectories become more concise. Concurrently, the frequency of reflection tokens exhibits an upward trend. This pattern is more distinct on the MATH dataset compared to GSM8K, which corresponds to the higher difficulty of the task. The combination of decreasing total length and increasing reflection frequency implies that the models utilize the structural patterns more efficiently rather than merely producing longer sequences.

Inference Behavior. We further analyze the inference behavior of TrajFusion from two perspectives: when reflection is triggered, and whether the quality of the initial prediction is preserved. Table 3 shows that reflection is much more likely to appear on instances where vanilla RFT fails, indicating that it is activated selectively rather than uniformly. Table 4 further shows that TrajFusion maintains comparable or slightly stronger first-pass accuracy than vanilla RFT across both backbones.

Table 3: **Conditional reflection frequency of TrajFusion.** Reflection is conditioned on whether vanilla RFT produces a correct solution.

Condition	MATH	College	DM
$P(\text{Reflect} \mid \text{RFT Correct})$	11.9	7.8	7.8
$P(\text{Reflect} \mid \text{RFT Wrong})$	47.2	51.1	20.9

Table 4: **First-pass accuracy comparison.** TrajFusion preserves or improves the quality of initial predictions relative to vanilla RFT.

Benchmark	LLaMA3-8B		DeepSeekMath-7B	
	RFT	TF@first	RFT	TF@first
MATH	48.5	50.9	56.2	57.9
College	28.9	31.3	40.1	39.9
DM	41.1	42.5	69.1	69.2

Taken together, these results suggest that TrajFusion learns an adaptive reasoning strategy that selectively invokes additional reasoning on harder problems without degrading the quality of initial predictions.

4.5 Token-Matched Comparison

TrajFusion constructs fused trajectories that include both incorrect and correct reasoning paths, resulting in longer training sequences than vanilla RFT under the same number of samples. To determine whether the observed gains arise from increased token exposure or from the supervision structure itself, we construct two controlled RFT variants. **RFT-Scaled** increases the total token budget by selecting multiple correct trajectories per problem, while **RFT-Long** increases sequence length by selecting longer correct trajectories without changing the number of samples. These variants isolate the effects of data scale and sequence length under the same training objective.

Across both backbones, increasing token exposure alone does not explain the improvements. RFT-Long, which only increases sequence length, fails to improve performance and can degrade it, indicating that longer correct trajectories do not provide additional useful supervision. RFT-Scaled yields moderate gains by increasing data volume, but remains consistently below TrajFusion despite using significantly more tokens. In contrast, TrajFusion achieves the best performance with fewer tokens than RFT-Scaled, demonstrating that its advantage stems from the structured integration of diverse reasoning trajectories rather than token quantity.

Table 5: **Token-matched comparison with controlled RFT variants.** RFT-Scaled increases data size, RFT-Long increases sequence length, while TrajFusion integrates incorrect trajectories.

Method	Tokens (M)	MATH	Coll.	Olym.	DM	Avg
<i>LLaMA3-8B</i>						
RFT	182.3	48.5	28.9	14.5	41.1	33.3
RFT-Scaled	366.3	49.8	30.2	16.9	42.4	34.8
RFT-Long	306.9	46.2	27.1	14.7	41.1	32.3
TrajFusion	298.0	51.4	31.6	20.2	43.7	36.7
<i>DeepSeekMath-7B</i>						
RFT	182.3	56.2	40.1	25.6	69.1	47.8
RFT-Scaled	366.3	57.4	40.5	25.9	69.1	48.2
RFT-Long	306.9	52.2	38.6	24.3	66.9	45.5
TrajFusion	298.0	59.1	41.2	27.0	69.6	49.2

These results support that the gains of TrajFusion arise from richer supervision signals rather than increased training scale.

4.6 Scaling to Long-form Reasoning Data

Setup. Recent advances in mathematical reasoning have been facilitated by large reasoning models (LRMs) (DeepSeek-AI, 2025; Moshkov et al., 2025; Yang et al., 2025), which train on long-form trajectories and operate with extended context windows. To examine whether TrajFusion scales to this setting, we apply it to Qwen2.5-Math-7B-base, extending its context window from 4,096 tokens to 32,768 using YaRN (Peng et al., 2023).

We construct long-form training data from Deepseek-R1-Distill-Qwen-7B (DeepSeek-AI, 2025) and apply TrajFusion by selecting diverse incorrect trajectories followed by a correct solution. Evaluation is performed on three challenging benchmarks: AIME24, AIME25, and MATH-500. AIME contains competition-style problems with small validation splits; we report both **Pass@1** and **Avg@16**. MATH-500 (Lightman et al., 2023) is larger and evaluated using Pass@1 only. To study the effect of context length, we also train 4K-context variants using data sampled from Qwen2.5-Math-7B-Instruct and compare vanilla RFT with TrajFusion.

Results. Table 6 shows that TrajFusion consistently outperforms vanilla RFT across all datasets and context lengths. In the 4K setting, TrajFusion improves Pass@1 from 16.7% to 23.3% on AIME24 and from 10.0% to 16.7% on AIME25. While the gains in Avg@16 are more moderate

Table 6: Performance of TrajFusion and vanilla RFT on AIME24, AIME25, and MATH-500 using **Qwen2.5-Math-7B-base**. Results are reported for both 4K and 32K context.

Method	AIME24	AIME25	MATH-500
	<i>P@1/Avg@16</i>	<i>P@1/Avg@16</i>	<i>P@1</i>
RFT(4K)	16.7 / 15.6	10.0 / 12.1	79.4
TrajFusion(4K)	23.3 / 15.6	16.7 / 13.8	82.6
RFT(32K)	40.0 / 36.7	26.7 / 27.9	88.4
TrajFusion(32K)	43.3 / 40.8	30.0 / 29.6	90.4

in this configuration, the substantial increase in Pass@1 suggests that TrajFusion effectively enhances the quality of top-ranked reasoning paths.

The performance advantage remains robust in the 32K setting utilizing long-form data. On AIME24, TrajFusion increases Pass@1 from 40.0% to 43.3% and Avg@16 from 36.7% to 40.8%. Similar improvements are observed on AIME25, and MATH-500 accuracy rises from 88.4% to 90.4%. These results indicate that TrajFusion scales effectively to extended context windows and complex reasoning trajectories.

5 Related Work

5.1 Data Construction for Math Reasoning

Prior work improves mathematical reasoning primarily by constructing high-quality solution data. Most approaches focus on expanding the coverage and diversity of *correct* reasoning trajectories, for example by modifying problem difficulty, reformulating questions, recombining solution components, or sampling multiple valid solution paths (Luo et al., 2023; Yu et al., 2023; Li et al., 2023; Tong et al., 2024; Guan et al., 2025). To ensure reliability, these pipelines typically apply strict correctness checks or rejection sampling, retaining only fully correct generations as supervision (Zelikman et al., 2022; Yuan et al., 2023b; Xu et al., 2024; Guan et al., 2025; Xiong et al., 2025). As a result, incorrect or partially incorrect trajectories are systematically excluded at the data level.

5.2 Negative Signals for Math Reasoning

Several approaches introduce negative signals to improve mathematical reasoning. Preference-based and reinforcement learning methods, such as DPO (Rafailov et al., 2023), GRPO (Shao et al., 2024), and their variants (Meng et al., 2024; Lu et al., 2024; Lai et al., 2024), leverage suboptimal or lower-reward samples to guide models away

from undesirable reasoning paths. NSR (Zhu et al., 2025) shows that exclusively optimizing negative samples in reinforcement learning yields significant gains. NFT (Chen et al., 2025) incorporates negative signals by modifying the supervised training loss. Self-correction approaches (An et al., 2023; Pan et al., 2025) introduce errors through multi-turn data synthesis, exposing models to flawed trajectories and their revisions. TrajFusion differs in that it operates within the RFT pipeline and directly utilizes discarded samples. It does not require preference modeling or loss modification, enabling single-stage learning from negative signals under standard next-token prediction.

6 Conclusion

In this work, we proposed TrajFusion, a lightweight fine-tuning method that leverages both correct and incorrect reasoning trajectories during rejection sampling fine-tuning. By structuring diverse failed completions and reflection prompts into adaptive fused trajectories, TrajFusion provides structured supervision by a compact trial-and-error reasoning path without requiring architectural changes or modifications to the training objective. Experiments across diverse math benchmarks and models show consistent gains over vanilla rejection fine-tuning, especially on challenging and long-form regimes. TrajFusion offers a simple yet effective way to scale mathematical reasoning through supervision.

7 Limitations

TrajFusion leverages supervision signals from problems where model sampling yields both correct and incorrect trajectories, as such cases provide the most informative reasoning variations. Consequently, the strength of the supervision signal depends on the presence of sufficient outcome diversity during sampling, while questions that are consistently solved correctly or incorrectly offer limited contrastive information, in which case TrajFusion naturally reduces to vanilla rejection sampling fine-tuning behavior.

References

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2023. [Learning from mistakes makes llm better reasoner](#). *CoRR*, abs/2310.20689.

- Huayu Chen, Kaiwen Zheng, Qinsheng Zhang, Ganqu Cui, Yin Cui, Haotian Ye, Tsung-Yi Lin, Ming-Yu Liu, Jun Zhu, and Haoxiang Wang. 2025. Bridging supervised learning and reinforcement learning in math reasoning. *arXiv preprint arXiv:2505.18116*.
- Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. **TheoremQA: A theorem-driven question answering dataset**. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- 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*.
- OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. <https://github.com/open-compass/opencompass>.
- DeepSeek-AI. 2025. **Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning**. *Preprint*, arXiv:2501.12948.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.
- Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng 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. *arXiv preprint arXiv:2402.14008*.
- Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, and 1 others. 2025. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *arXiv preprint arXiv:2504.11456*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. **Measuring mathematical problem solving with the math dataset**. *CoRR*, abs/2103.03874.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangu Peng, and Jiaya Jia. 2024. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*.
- Chengpeng Li, Zheng Yuan, Hongyi Yuan, Guanting Dong, Keming Lu, Jiancan Wu, Chuanqi Tan, Xiang Wang, and Chang Zhou. 2023. Mugglemath: Assessing the impact of query and response augmentation on math reasoning. *arXiv preprint arXiv:2310.05506*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*.
- Haoxiong Liu, Yifan Zhang, Yifan Luo, and Andrew C Yao. 2025. Augmenting math word problems via iterative question composing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24605–24613.
- Zimu Lu, Aojun Zhou, Ke Wang, Houxing Ren, Weikang Shi, Juntong Pan, Mingjie Zhan, and Hongsheng Li. 2024. Step-controlled dpo: Leveraging stepwise error for enhanced mathematical reasoning. *arXiv preprint arXiv:2407.00782*.
- Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.
- Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.
- Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. 2025. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *arXiv preprint arXiv:2504.16891*.
- Zhuoshi Pan, Yu Li, Honglin Lin, Qizhi Pei, Zinan Tang, Wei Wu, Chenlin Ming, H. Vicky Zhao, Conghui He, and Lijun Wu. 2025. **Lemma: Learning from errors for mathematical advancement in LLMs**. *CoRR*, abs/2503.17439.
- Qizhi Pei, Lijun Wu, Zhuoshi Pan, Yu Li, Honglin Lin, Chenlin Ming, Xin Gao, Conghui He, and Rui Yan. 2025. Mathfusion: Enhancing mathematical problem-solving of llm through instruction fusion. *arXiv preprint arXiv:2503.16212*.

- Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.
- 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. *Advances in neural information processing systems*, 36:53728–53741.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506.
- David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. [Analysing mathematical reasoning abilities of neural models](#). In *International Conference on Learning Representations*.
- 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*.
- Zhengyang Tang, Xingxing Zhang, Benyou Wang, and Furu Wei. 2024. Mathscales: Scaling instruction tuning for mathematical reasoning. *arXiv preprint arXiv:2403.02884*.
- Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024. [Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving](#). In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). *CoRR*, abs/2201.11903.
- Wei Xiong, Jiarui Yao, Yuhui Xu, Bo Pang, Lei Wang, Doyen Sahoo, Junnan Li, Nan Jiang, Tong Zhang, Caiming Xiong, and 1 others. 2025. A minimalist approach to llm reasoning: from rejection sampling to reinforce. *arXiv preprint arXiv:2504.11343*.
- Yifan Xu, Xiao Liu, Xinghan Liu, Zhenyu Hou, Yueyan Li, Xiaohan Zhang, Zihan Wang, Aohan Zeng, Zhengxiao Du, Wenyi Zhao, and 1 others. 2024. Chatglm-math: Improving math problem-solving in large language models with a self-critique pipeline. *arXiv preprint arXiv:2404.02893*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. [Meta-math: Bootstrap your own mathematical questions for large language models](#). *CoRR*, abs/2309.12284.
- Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023a. Rrhf: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems*, 36:10935–10950.
- Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023b. [Scaling relationship on learning mathematical reasoning with large language models](#). *Preprint*, arXiv:2308.01825.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.
- Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024. [Swift: a scalable lightweight infrastructure for fine-tuning](#). *Preprint*, arXiv:2408.05517.
- Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. 2025. The surprising effectiveness of negative reinforcement in llm reasoning. *arXiv preprint arXiv:2506.01347*.

A Experimental Details

A.1 Sampling Setup

We adopt a diverse sampling strategy to expose the model to both successful and unsuccessful reasoning paths.

Sampling Strategy. For each input query, we generate 16 independent completions using nucleus sampling ($\text{top-}p = 0.95$) with a temperature of 0.7. This configuration balances diversity and coherence, encouraging exploration of multiple reasoning trajectories without producing excessively random outputs.

Prompt Format. All completions are generated using a fixed instruction-style prompt:

```
Please reason step by step and put your answer within \boxed{}
```

This encourages models to follow a chain-of-thought reasoning process and produce a clearly marked final answer suitable for automatic extraction.

Output Filtering. To ensure semantically valid supervision signals, we filter generated responses to remove malformed outputs. Specifically, we discard completions that:

- exhibit excessive repetition;
- switch languages mid-response (e.g., mixing English and Chinese);
- lack meaningful reasoning steps or a final boxed answer.

This step ensures that the training data remains pedagogically useful.

Answer Extraction and Verification. We extract the final boxed or inline answer from each completion using simple pattern matching heuristics. To assess correctness, we apply the automatic math verifier from opencompass (Contributors, 2023), which executes symbolic evaluation of the extracted answer against the ground truth. Each completion is labeled as *correct* or *incorrect* based on exact match criteria.

Model-Specific Configurations. Sampling is conducted using two teacher models: **Qwen2.5-Math-7B-Instruct** (4K context) and **DeepSeek-R1-Distill-Qwen-7B** (16K context). Each model

uses its default maximum context length and the same sampling settings above. Queries for which no correct completions are found are excluded from training.

A.2 Trajectory Fusion

Construction. For each query, we identify the set of incorrect responses generated during rejection sampling. These responses are grouped by their final extracted values, and we select the shortest representative from each group to prioritize concise reasoning paths with minimal redundancy.

Sequence Formatting. A complete assistant reply is constructed by interleaving the selected incorrect responses with fixed reflective prompts, simulating a trial-and-error reasoning process. The sequence concludes with a correct solution followed by a positive feedback phrase and the final answer. The trajectory is encapsulated within a thinking block to provide structured supervision:

```
<think>
[Incorrect Response 1]
[Reflective Comment]
...
[Correct Response]
[Success Phrase]
</think>
The answer is \boxed{[Answer]}.
```

If no eligible incorrect completions are found, the sequence defaults to a standard format containing only the correct response within the `<think>` block, omitting the error and reflection components.

A.3 Training Setup

Framework and Resources. We perform training using the `ms-swift` framework with DeepSpeed ZeRO-1 optimization. All models are trained on $8 \times A100$ 80GB GPUs using bf16 mixed-precision and input sequence packing to improve efficiency.

Hyperparameters. Table 7 provides the detailed training configurations for each base model. We evaluate models at fixed data intervals within the first epoch to monitor performance dynamics (Figure 3).

Table 7: Training configurations for different base models and data scales.

Base Model	Teacher Model	Learning Rate	Length	Context	#Samples
LLaMA3-8B-base	Qwen2.5-Math-7B-instruct	5e-5	4096	Standard	15K/100K
DeepSeekMath-7B-base	Qwen2.5-Math-7B-instruct	5e-5	4096	Standard	15K/100K
Qwen2.5-Math-7B-base	Qwen2.5-Math-7B-instruct	2e-5	4096	Standard	15K
Qwen2.5-Math-7B-base	DeepSeek-R1-Distill-Qwen-7B	2e-5	32768	Extended	100K

B Prompts and Templates

B.1 Dialogue Templates

To ensure the models correctly interpret the structured reasoning tasks, we employ native dialogue formats for each model family. The following templates illustrate the precise placement of system instructions, user queries, and assistant responses.

Qwen2.5-Math Template

```
<|im_start|>system
Please reason step by step, and put your
  final answer within \boxed{ }.
<|im_end|>
<|im_start|>user
{input}
<|im_end|>
<|im_start|>assistant
{output}
```

LLaMA3 Template

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
Please reason step by step, and put your
  final answer within \boxed{ }.
<|eot_id|>
<|start_header_id|>user<|end_header_id|>
{input}
<|eot_id|>
<|start_header_id|>assistant
<|end_header_id|>
{output}
<|eot_id|>
```

DeepSeek-Math Template

```
User: {input}
Please reason step by step, and put your
  final answer within \boxed{ }.

Assistant: {output}
```

B.2 Reflective Prompt Repository

To internalize the trial-and-error reasoning process, TrajFusion utilizes a curated set of transition

phrases. These prompts are inserted between reasoning trajectories to signify error recognition and subsequent correction. Table 8 lists the complete repository of phrases used in our experiments.

B.3 A Example of TrajFusion

Below we provide a simple TrajFusion example.

Example of TrajFusion

Problem: Determine $(r+s)(s+t)(t+r)$, given that r, s, t are the real roots of $x^3 + 9x^2 - 9x - 8$.

TrajFusion Solution:

Given the polynomial $P(x) = x^3 + 9x^2 - 9x - 8$ with roots r, s, t , we need to determine the value of $(r+s)(s+t)(t+r)$. By Vieta's formulas, we know: $r + s + t = -9$, $rs + rt + st = -9$, $rst = 8$. We first try to expand the expression as $(r+s)(s+t)(t+r) = rs + rt + st + 2rst$, and substituting the values we get $-9 + 2 \cdot 8 = 7$.

Wait... that doesn't seem right. Let's try again.

Given the polynomial $P(x) = x^3 + 9x^2 - 9x - 8$ with roots r, s, t , we need to find the value of $(r+s)(s+t)(t+r)$. By Vieta's formulas, we know: $r + s + t = -9$, $rs + rt + st = -9$, $rst = 8$. We start by expressing $(r+s)(s+t)(t+r)$ in a different form: $(r+s)(s+t)(t+r) = (r+s+t)(rs + rt + st) - rst$. Substituting the values from Vieta's formulas, we get: $(-9)(-9) - 8 = 81 - 8 = 73$. Thus, the value of $(r+s)(s+t)(t+r)$ is $\boxed{73}$. **Ah, that makes sense now.**

B.4 Evaluation Protocol

Setup. All evaluations were conducted using the official Qwen2.5-Math toolkit, with inference executed through the same vLLM engine and prompt templates as used in training. Decoding was configured with temperature 0.6 and top- p 0.95, encouraging diverse reasoning while maintaining output coherence.

Metrics. For competition-style datasets such as AIME24 and AIME25, we report both **Pass@1** and **Avg@16** using 16 sampled completions per query. For all other datasets, we evaluate using a single generation and report Pass@1. Table 9 summarizes these settings.

Benchmarks. We evaluate model performance across these mathematical reasoning benchmarks that vary in domain, difficulty, and format:

Table 8: Complete repository of reflection and feedback phrases used in TrajFusion.

Category	Candidate Phrases (Randomly Sampled)
Error Recognition	"Wait... that doesn't seem right. Let's try again." "Hmm, maybe we're missing something. Think carefully." "That doesn't add up. Let's give it another go." "Hold on. That might not be correct. Try again."
Success Confirmation	"That makes sense now." "Yes, this is correct." "Ah, that makes sense now."
Certainty (Optional)	"I'm absolutely certain this is correct." "No doubt about it, this answer is right." "I fully trust this result."

Table 9: Evaluation Settings by Dataset

Dataset	Samples per Query	Reported Metrics
AIME24	16	Pass@1, Avg@16
AIME25	16	Pass@1, Avg@16
All others	1	Pass@1

- **GSM8K** contains 8,792 grade-school math word problems focused on multi-step arithmetic reasoning. Each question typically requires 2–8 steps to solve. We use the official test split of 1,319 examples.
- **MATH** is a competition-style dataset of 12,500 problems drawn from high school mathematics contests. It spans seven categories—ranging from algebra to geometry—and includes problems at five difficulty levels. We evaluate on the 5,000-problem test set.
- **CollegeMath** includes 2,818 curated problems sourced from undergraduate math textbooks. It covers topics such as algebra, calculus, probability, linear algebra, and differential equations, targeting college-level mathematical proficiency.
- **DeepMind Mathematics** consists of 1,000 short-form math problems designed to evaluate symbolic and numerical reasoning. The dataset covers core areas like arithmetic, algebra, calculus, and basic probability.
- **OlympiadBench-Math** is a benchmark of 675 Olympiad-style problems aimed at evaluating

advanced mathematical reasoning. We use only the English text-only subset in our experiments.

- **TheoremQA** is a theorem-centric question-answering dataset built around 350 mathematical and scientific theorems. It includes 800 problems spanning domains such as mathematics, physics, computer science, and engineering, designed to test a model's ability to apply domain-specific knowledge.
- **MATH-500** is a carefully curated subset of 500 math problems drawn from diverse domains such as algebra, probability, geometry, and trigonometry. These problems are designed to assess a model's ability to engage with structured, multi-step reasoning under compact evaluation settings.
- **AIME24** consists of 30 problems from the 2024 American Invitational Mathematics Examination, an upper-tier contest for top AMC performers. Each question demands intensive algebraic, geometric, or combinatorial reasoning, and requires an integer answer between 0 and 999.
- **AIME25** is the corresponding AIME-style test from 2025, offering a separate but structurally similar set of 30 high-difficulty problems to reduce potential overlap and memorization.

C Case Study

We present some representative cases from the test set to qualitatively analyze the behavior of DeepSeek-Math-7B-TrajFusion.

A Case from TrajFusion-DSMath

Problem: Simplify

$$\frac{\sec x}{\sin x} - \frac{\sin x}{\cos x}.$$

TrajFusion-DSMath Solution:

<think> To simplify the expression $\frac{\sec x}{\sin x} - \frac{\sin x}{\cos x}$, we start by expressing $\sec x$ and $\sin x$ in terms of $\cos x$ and $\sin x$:

$$\sec x = \frac{1}{\cos x} \quad \text{and} \quad \sin x = \sin x.$$

Substituting these into the expression, we get:

$$\frac{\sec x}{\sin x} = \frac{\frac{1}{\cos x}}{\sin x} = \frac{1}{\cos x \sin x} \quad \text{and} \quad \frac{\sin x}{\cos x} = \frac{\sin x}{\cos x}.$$

So the expression becomes:

$$\frac{1}{\cos x \sin x} - \frac{\sin x}{\cos x}.$$

To combine these fractions, we need a common denominator. The common denominator for $\cos x \sin x$ and $\cos x$ is $\cos x \sin x$. Rewriting the second fraction with this common denominator, we get:

$$\frac{\sin x}{\cos x} = \frac{\sin x \sin x}{\cos x \sin x} = \frac{\sin^2 x}{\cos x \sin x}.$$

Now we can combine the fractions:

$$\frac{1}{\cos x \sin x} - \frac{\sin^2 x}{\cos x \sin x} = \frac{1 - \sin^2 x}{\cos x \sin x}.$$

We recognize that $1 - \sin^2 x$ is a difference of squares, which can be factored as:

$$1 - \sin^2 x = (\cos x + \sin x)(\cos x - \sin x).$$

So the expression becomes:

$$\frac{(\cos x + \sin x)(\cos x - \sin x)}{\cos x \sin x}.$$

We can split this into two fractions:

$$\frac{\cos x + \sin x}{\cos x} \cdot \frac{\cos x - \sin x}{\sin x}.$$

This simplifies to:

$$\left(1 + \frac{\sin x}{\cos x}\right) \left(\frac{\cos x}{\sin x} - 1\right).$$

We know that $\frac{\sin x}{\cos x} = \tan x$ and $\frac{\cos x}{\sin x} = \cot x$, so the expression becomes:

$$(1 + \tan x)(\cot x - 1).$$

Using the identity $\cot x = \frac{1}{\tan x}$, we can rewrite the expression as:

$$(1 + \tan x) \left(\frac{1}{\tan x} - 1\right) = (1 + \tan x) \cdot \frac{1 - \tan x}{\tan x} = \frac{(1 + \tan x)(1 - \tan x)}{\tan x} = \frac{1 - \tan^2 x}{\tan x}.$$

We know that $1 - \tan^2 x = \sec^2 x - \tan^2 x = \sec^2 x - \frac{\sin^2 x}{\cos^2 x} = \frac{\cos^2 x}{\cos^2 x} - \frac{\sin^2 x}{\cos^2 x} = \frac{\cos^2 x - \sin^2 x}{\cos^2 x} = \frac{\cos 2x}{\cos^2 x}$, so the expression becomes:

$$\frac{\frac{\cos 2x}{\cos^2 x}}{\tan x} = \frac{\cos 2x}{\cos^2 x} \cdot \frac{\cos x}{\sin x} = \frac{\cos 2x \cos x}{\cos^2 x \sin x} = \frac{\cos 2x}{\cos x \sin x} = \frac{2 \cos^2 x - 1}{\cos x \sin x} = \frac{2 \cos x}{\sin x} - \frac{1}{\cos x \sin x} = 2 \cot x - \csc x.$$

Therefore, the simplified form of the expression is:

$$\boxed{2 \cot x - \csc x}.$$

Wait... that doesn't seem right. Let's try again.

To simplify the expression $\frac{\sec x}{\sin x} - \frac{\sin x}{\cos x}$, we start by expressing $\sec x$ and $\sin x$ in terms of $\cos x$ and $\sin x$:

$$\sec x = \frac{1}{\cos x} \quad \text{and} \quad \sin x = \sin x.$$

Substituting these into the expression, we get:

$$\frac{\sec x}{\sin x} = \frac{\frac{1}{\cos x}}{\sin x} = \frac{1}{\cos x \sin x} \quad \text{and} \quad \frac{\sin x}{\cos x} = \frac{\sin x}{\cos x}.$$

So the expression becomes:

$$\frac{1}{\cos x \sin x} - \frac{\sin x}{\cos x}.$$

To combine these fractions, we need a common denominator. The common denominator for $\cos x \sin x$ and $\cos x$ is $\cos x \sin x$. Rewriting the second fraction with this common denominator, we get:

$$\frac{\sin x}{\cos x} = \frac{\sin x \sin x}{\cos x \sin x} = \frac{\sin^2 x}{\cos x \sin x}.$$

Now we can combine the fractions:

$$\frac{1}{\cos x \sin x} - \frac{\sin^2 x}{\cos x \sin x} = \frac{1 - \sin^2 x}{\cos x \sin x}.$$

We know from the Pythagorean identity that $1 - \sin^2 x = \cos^2 x$. Substituting this into the expression, we get:

$$\frac{\cos^2 x}{\cos x \sin x} = \frac{\cos x}{\sin x}.$$

The expression $\frac{\cos x}{\sin x}$ is the definition of $\cot x$. Therefore, the simplified form of the original expression is:

$$\boxed{\cot x}.$$

That makes sense now.

</think>The answer is

$$\boxed{\cot x}.$$

A Case from TrajFusion-DSMath-7B

Problem: If $\frac{\sqrt{2x}}{\sqrt{3x-1}} = \frac{3}{2}$, solve for x . Express your answer in simplest fractional form.

TrajFusion-DSMath Solution:

<think> To solve the equation $\frac{\sqrt{2x}}{\sqrt{3x-1}} = \frac{3}{2}$, we start by eliminating the square roots by squaring both sides of the equation:

$$\left(\frac{\sqrt{2x}}{\sqrt{3x-1}} \right)^2 = \left(\frac{3}{2} \right)^2$$

This simplifies to:

$$\frac{2x}{3x-1} = \frac{9}{4}$$

Next, we eliminate the fractions by cross-multiplying:

$$4 \cdot 2x = 9 \cdot (3x - 1)$$

This simplifies to:

$$8x = 27x - 9$$

To isolate x , we subtract $8x$ from both sides:

$$0 = 19x - 9$$

Then, we add 9 to both sides:

$$9 = 19x$$

Finally, we divide both sides by 19:

$$x = \frac{9}{19}$$

Thus, the solution is:

$$\frac{9}{19}$$

</think>

The answer is

$$\frac{9}{19}$$