ReFLAIR: Enhancing Multimodal Reasoning via Structured Reflection and Reward-Guided Learning

Jiazhou Ji[*] and **Xinru Lu**[†]
Shanghai Jiao Tong University
{jjz1011,later1}@sjtu.edu.cn

Abstract

Large models can achieve higher performance on complex problems through iterative self-reflection. Yet when reflection is uncontrolled, it often leads to longer outputs, higher inference cost, and an increased risk of hallucination. Existing training methods rarely address this trade off. We introduce **ReFLAIR**, a unified framework that teaches multimodal large models to perform structured reflection via an explicit <think><re-think><answer> format and hybrid reward learning. ReFLAIR begins with supervised cold start training on the ReFLAIR-cold dataset of curated multimodal reasoning trajectories, and then trains a *Reflection Quality Scorer* (RQS) to quantify the utility of rethinking steps. A modified grouped relative policy optimization algorithm optimizes a hybrid reward that combines answer correctness, structural fidelity, reflection utility, and sample difficulty. Evaluated on challenging mathematical benchmarks including MathVista, MathVerse, MM-Math and GSM8K, ReFLAIR yields improvements up to +12.2% absolute accuracy, produces higher quality reflective traces, and reduces harmful or redundant revisions. An adaptive test time reflection scheduler further reduces inference cost by nearly 25% while maintaining or improving accuracy. These results demonstrate that structured, reward guided reflection offers a scalable pathway to more reliable and interpretable reasoning in multimodal models.

1 Introduction

Large language models (LLMs) have achieved remarkable success in complex reasoning tasks, aided by techniques like chain-of-thought (CoT) prompting (Wei et al., 2022), program-aided reasoning, and reinforcement learning (OpenAI, 2024; DeepSeek-AI et al., 2025). These advances enable near-human or superhuman accuracy on benchmarks such as GSM8K and AIME. Multimodal large models (LMMs) have further extended this capability to vision-language tasks, reasoning over inputs like charts, equations, and diagrams. Recent surveys and benchmarks (Fu et al., 2023; Sun et al., 2024; Lu et al., 2024; Huang et al., 2024a; Masry et al., 2021; Wang et al., 2024b) highlight both rapid progress and the continued difficulty of handling complex visual reasoning. In parallel, reflection—where models iteratively evaluate and revise their own outputs—has emerged as a promising path for improving reasoning depth and robustness (Zelikman et al., 2022; Madaan et al., 2023; Ranaldi and Freitas, 2024; Shinn et al., 2023; Yan et al., 2024; Chen et al., 2025).

However, reflection remains underdeveloped in many state-of-the-art systems. For example, **Vision-R1** (Huang et al., 2025) adopts a <think><answer> structure and achieves strong results, while also noting the prevalence of reflective tokens in its training corpus. Yet, its final model does not explicitly reinforce reflection, leaving this capability underexploited. Conversely, **Satori** (Shen et al., 2025) introduces a more sophisticated Chain-of-Action-Thought format with custom tokens that promote self-reflection. Despite this, it lacks fine-grained control over the quality and utility of each reflective step, often generating redundant or unfocused reasoning. These limitations point to a gap: existing approaches either omit structured reflection or apply it without targeted guidance, leading to inefficiencies or hallucinations.

To address these limitations, we present **ReFLAIR**, a unified training framework that treats structured reflection as a first-class component of reasoning. The core idea is to scaffold multi-stage cognition with an explicit reflection structure, optimized through a combination of supervised learning and reinforcement learning. Supervised train-

[*]Corresponding Author.
[†]Equal Contribution.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 25401–25413
November 4-9, 2025 ©2025 Association for Computational Linguistics

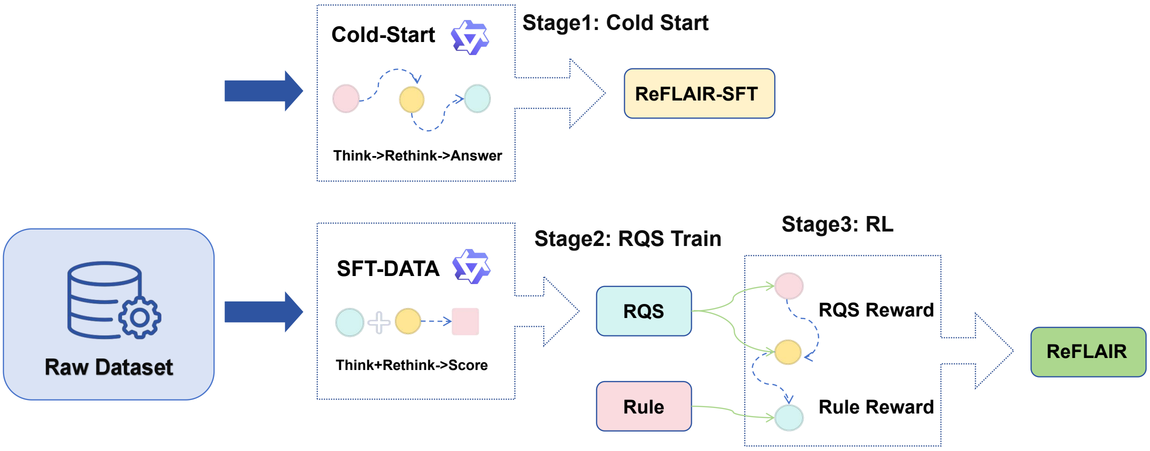

Figure 1: Overview of **ReFLAIR**, which integrates supervised cold-start training, reflection quality scoring, and reinforcement learning with hybrid rewards to foster reliable and structured reflective reasoning in MLLMs.

ing on the ReFLAIR-cold dataset provides initial reflective supervision by exposing models to diverse reasoning trajectories that cover exploratory and corrective revision. Reinforcement learning then enhances this capability through a hybrid reward scheme that integrates correctness, structural adherence, and a learned *Reflection Quality Scorer* (RQS). RQS is trained to estimate the value added by reflection steps using scalar annotations derived from reference trajectories, which enables more precise control over the utility of reflection. To ensure stability during optimization, the framework incorporates recent advances in preference-based reinforcement learning such as GRPO and DAPO (DeepSeek-AI et al., 2025; Yu et al., 2025). This training paradigm encourages models to produce reflective steps that are not merely verbose but demonstrably beneficial for downstream correctness. Illustrative examples of the reflection structure and generated trajectories are provided in Appendix A.

We evaluate ReFLAIR primarily on mathematical reasoning benchmarks. Our experimental suite encompasses multimodal and textual math datasets used across the paper, and the empirical study focuses on (i) end-to-end improvements in accuracy, (ii) ablation analyses that quantify the contribution of cold-start supervised trajectories, the <re-think> structural signal, and RQS supervision, (iii) transfer experiments showing that the reflection-aware training provides gains when applied to smaller textual LLMs, and (iv) analyses of test-time reflection scaling, where an adaptive scheduling mechanism reduces inference cost while maintaining or improving accuracy. Across these dimensions, ReFLAIR consistently produces more accurate and interpretable reasoning traces compared to baselines that lack targeted reflective supervision.

Our contributions are threefold. First, we introduce ReFLAIR, a training framework that explicitly scaffolds multi-stage reflection and aligns it with task utility. Second, we develop a reflection-aware reward scheme that integrates correctness, format adherence, and a learned reflection quality signal within a stabilized preference-optimization procedure. Third, we present comprehensive empirical evidence on mathematical reasoning benchmarks, including ablations, transfer studies to textual LLMs, and test-time scaling analyses, demonstrating that structured reflection yields reliable gains in both performance and interpretability. Please refer to the attached source code for further details[1].

2 Related Work

2.1 Prompting and Chain-of-Thought Reasoning

Large language models benefit substantially from prompting strategies that elicit structured reasoning. Chain-of-Thought prompting (Wei et al., 2022) enables models to generate intermediate steps, improving accuracy on complex problems. Building on this idea, Self-Consistency (Zelik-

[1] `https://github.com/jjz1011/ReFLAIR`

man et al., 2022), Least-to-Most prompting (Zhang et al., 2022), and Tree-of-Thought reasoning propose different ways of sampling or decomposing reasoning trajectories. Beyond prompting, reasoning patterns have been distilled into smaller models through instruction tuning (Madaan et al., 2023), making structured reasoning more broadly deployable. These methods achieve strong performance on benchmarks such as GSM8K (Cobbe et al., 2021), SVAMP, and MATH (Hendrycks et al., 2021a).

2.2 Reinforcement Learning for Long-form Reasoning

Reinforcement learning (RL) provides another paradigm for improving reasoning ability in large models. OpenAI's o1 model (OpenAI, 2024) and systems like DeepSeek-R1 apply RL to encourage models to internalize reasoning strategies during generation. RL has also been widely used for preference alignment, with RLHF (Stiennon et al., 2020) as a canonical example, and extended to adaptive computation settings (Guo et al., 2025). Recent advances explore alternative formulations, including DAPO (Yu et al., 2025), GRPO, and multi-head feedback mechanisms (Li et al., 2025), showing the flexibility of RL in aligning long-form reasoning with desirable outcomes.

2.3 Reflection and Iterative Self-Improvement

An emerging line of work focuses on enabling models to critique and revise their own outputs. Early approaches such as STaR (Zelikman et al., 2022) and Reflexion (Shinn et al., 2023) demonstrate the benefits of leveraging model-generated feedback to improve reasoning. Later methods like Self-Refine (Ranaldi and Freitas, 2024), Hindsight (Li et al., 2024b), and Mirror (Yan et al., 2024) refine this process through iterative revision, multi-path validation, or multiple-perspective reflection. These approaches vary in whether they rely on heuristic rules, oracle supervision, or internally generated critiques, but collectively highlight the promise of self-feedback for reasoning improvement.

2.4 Benchmarks and Evaluation

The evaluation of reasoning has expanded beyond simple answer accuracy toward more diagnostic settings. Classical benchmarks such as GSM8K, MATH (Hendrycks et al., 2021a), and AQuA remain central, while newer multimodal and domain-specific datasets such as MathVista (Lu et al., 2024), ChartQA (Masry et al., 2021), and MathVerse (Huang et al., 2024a) test mathematical and visual reasoning. Recent resources specifically target reflection and revision, including Mirror (Yan et al., 2024) and LR^{2}Bench (Chen et al., 2025), which introduce datasets such as MathReflex-10K and ReflectQA to measure iterative self-improvement across math and multimodal domains. Together, these benchmarks provide a rigorous basis for evaluating both chain-of-thought reasoning and reflective refinement.

3 Method

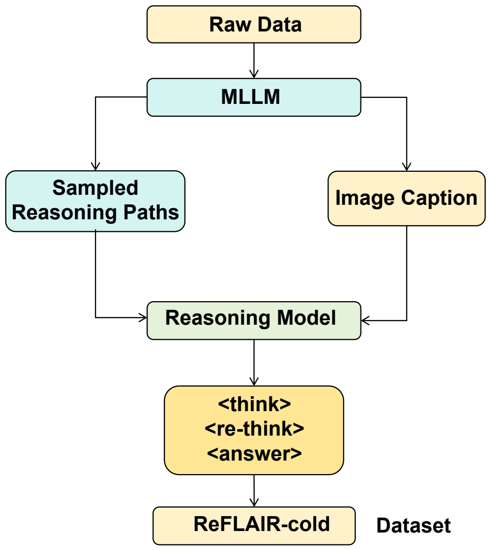

Figure 2: Pipeline for generating cold-start data.

We present **ReFLAIR**, a training framework designed to enhance the reflective reasoning capabilities of multimodal large language models (MLLMs). As illustrated in Figure 1, ReFLAIR consists of three key components: (1) cold-start supervised fine-tuning (SFT) on structured reflective reasoning trajectories, (2) a learned Reflection Quality Scorer (RQS) that estimates the value of model self-reflection, and (3) reinforcement learning with a hybrid reward function that encourages innovative and meaningful reflection. Each component is detailed below.

3.1 Overview of ReFLAIR

This cold-start stage is not merely preparatory — it endows the model with an explicit cognitive scaffold, teaching it to internalize the semantics of `<think>`, `<re-think>`, and `<answer>` as mean-

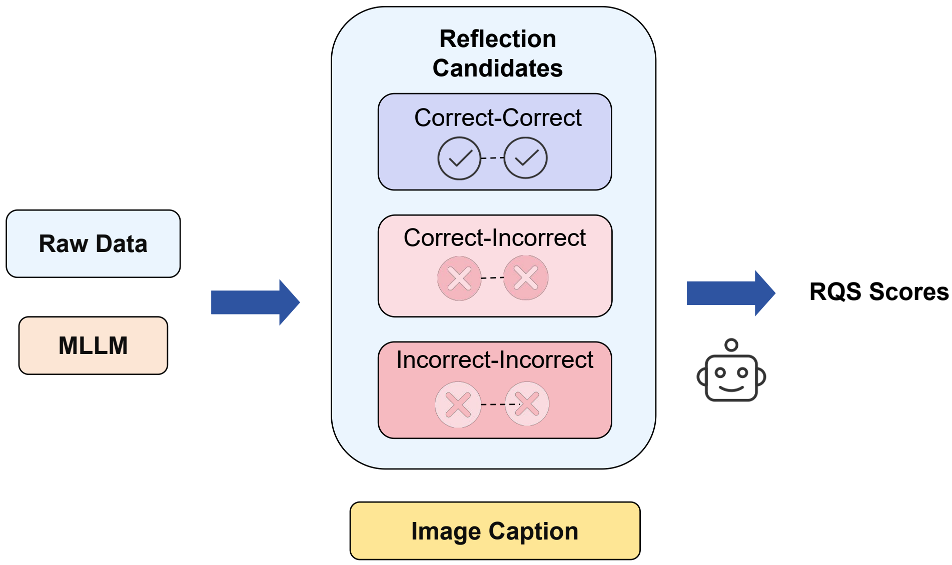

Figure 3: Pipeline for generating training data for RQS.

ingful reasoning markers. Through exposure to carefully curated, high-quality reflective trajectories, ReFLAIR-SFT learns to treat reflection as a first-class citizen in reasoning — not only generating answers, but understanding how and why an answer emerges through multi-step introspection.

To further optimize the model, we apply reinforcement learning to ReFLAIR-SFT, yielding the final **ReFLAIR** model. The reward function is carefully designed as a hybrid of model-based and rule-based signals. The model-based component is derived from the Reflective Quality Scorer (RQS), which provides a learned, high-level assessment of reasoning quality. The rule-based components ensure fidelity to task goals by explicitly scoring the correctness of the final answer and structural adherence to the <think><re-think><answer> format. This dual mechanism encourages the model to align with both latent human-like reasoning patterns and explicit task constraints, enabling ReFLAIR to generate responses that are not only accurate, but also cognitively coherent and structurally principled.

3.2 Cold-Start Training with Reflective Paths

To initialize ReFLAIR with a strong inductive bias toward reflective reasoning, we construct a multimodal dataset annotated with structured <think><re-think><answer> sequences. As illustrated in Figure 2, we begin with math-focused image–question–answer datasets and augment them through the following processes:

Exploring alternative reasoning strategies. For each image–question pair, we use a pretrained multimodal model to sample multiple reasoning paths, filtering them based on answer correctness. In parallel, we generate a detailed caption of the image to support grounded reasoning. Given the question, caption, and the pool of sampled reasoning paths, we prompt DeepSeek-R1 to select a pair of correct solutions that differ significantly in approach, as well as a pair consisting of one correct and one incorrect solution with clear logical divergence. These selected pairs are then refined by DeepSeek-R1 into coherent <think><re-think><answer> sequences. The correct–correct pairs model reflection as exploration of alternative valid strategies, while the incorrect–correct pairs promote reflective shifts from flawed logic to sound reasoning.

Step-level error correction. To further model introspective reasoning, we use a multimodal model to generate single sampled solutions for each image–question pair and obtain corresponding image captions. These solutions, which may contain suboptimal reasoning, are paired with their associated question and caption and passed to DeepSeek-R1. The model identifies critical flaws in the reasoning chain and generates a revised reflective path, which is only retained if the final answer is correct. The resulting <think><re-think><answer> sequences are fluently rewritten by DeepSeek-R1. This process strengthens the model's ability to both reinforce sound reasoning through focused reflection and recover from errors by re-evaluating key steps.

The resulting dataset, **ReFLAIR-cold**, consists of high-quality reflective trajectories spanning both strategic exploration and targeted error correction. It is used to train the initial ReFLAIR-SFT model, equipping it with the ability to produce structured, cognitively motivated reasoning from the outset.

3.3 Reflection Quality Scorer and Its Training

The **Reflection Quality Scorer (RQS)** is designed to assess whether a model's reflection genuinely improves the reasoning process. High-quality reflections either correct flawed logic or introduce effective new reasoning strategies. In contrast, shallow paraphrasing, repetition, or logically misleading modifications are penalized.

We define a **scoring rubric** to categorize reflection quality into four levels:

Fundamental Shift (0.8–1.0): reflections that transform an incorrect or suboptimal solution into a correct and structurally different one.

Insightful Revision (0.3–0.7): reflections that revise key reasoning steps or provide useful insights, while largely maintaining the original strategy.

Surface Edit (0.0 to –0.5): reflections that result in only superficial or stylistic changes, without improving reasoning quality.

Harmful Edit (–2.0): reflections that reinforce errors or introduce new mistakes into the reasoning process.

To construct training data for RQS, we build a dedicated dataset named **ReFLAIR-RQS** (see Figure 3). For each image–question pair, we sample multiple reasoning paths and answers using a multimodal model and generate a detailed image caption. From these, we form reflection candidates across three categories: (1) correct–correct (divergent yet valid strategies), (2) correct–incorrect, and (3) incorrect–incorrect. Given the question, caption, candidate reasoning paths, and the rubric, we prompt DeepSeek-R1 to generate a reflection quality score and a natural language explanation.

We then fine-tune a small language model using the resulting **ReFLAIR-RQS** dataset. Training is conducted with both *regression loss* (to predict scalar reflection quality scores) and *pairwise ranking loss* (to enforce relative quality ordering). The model is supervised with both scores and explanations. Once trained, RQS is frozen and used as a model-based reward component during reinforcement learning, evaluating the introspective quality of the <re-think> step in generated sequences.

3.4 Reinforcement Learning with Structured Rewards

To further enhance reflection-aware reasoning, we apply a structured reinforcement learning approach based on GRPO (DeepSeek-AI et al., 2025), extending the DAPO framework (Yu et al., 2025) to integrate reflective reasoning signals and learned rewards.

The total reward $\mathcal{R}$ assigned to a sampled reasoning trajectory is defined as:

$$\mathcal{R} = \lambda_1 \cdot \mathcal{R}_{\text{acc}} + \lambda_2 \cdot \sqrt{\mathcal{R}_{\text{rqs}} \cdot \mathcal{D}} + \lambda_3 \cdot \mathcal{R}_{\text{fmt}}$$

where $\mathcal{R}_{\text{acc}} \in \{0, 1\}$ indicates the correctness of the final answer, $\mathcal{R}_{\text{rqs}} \in [-2, 1]$ is the reflection quality score given by the RQS model, and $\mathcal{R}_{\text{fmt}} \in \{0, 1\}$ is a structural reward indicating whether the output adheres to the <think><re-think><answer> format. The difficulty score $\mathcal{D} = 1 - \hat{p} + b$ adjusts for sample difficulty, where $\hat{p}$ is the empirical accuracy of the base model on the same instance across multiple samples, and b is a small positive bias term for stability.

This reward formulation encourages thoughtful reflection on difficult problems, while minimizing redundant or unmotivated elaboration on easier ones. We adopt GRPO to optimize grouped response candidates. Following DAPO (Yu et al., 2025), we increase the upper clipping bound (`high`) to promote exploration and avoid premature convergence to suboptimal tokens.

4 Experiments

4.1 Experiment Settings

Datasets and Benchmarks. We construct three reflective reasoning datasets, named **ReFLAIR-cold**, **ReFLAIR-RQS** and **ReFLAIR-RL**, by selecting and transforming examples drawn from four large-scale sources: MathVision (Wang et al., 2024a, 2025), PolyMath (Gupta et al., 2024), OpenMathInstruct-2 (Toshniwal et al., 2024), and NuminaMath-CoT (Li et al., 2024a). The two datasets **ReFLAIR-cold** and **ReFLAIR-RQS** are constructed following the pipelines illustrated in Figure 2 and Figure 3 respectively, incorporating format restructuring, quality filtering, and reflection prompting. We evaluate our models on five mathematical reasoning benchmarks: MathVista (Liu et al., 2024), MathVerse (Huang et al., 2024b), MM-Math (Wang et al., 2024c), GSM8K (Cobbe et al., 2021), and MATH500 (Hendrycks et al., 2021b).

Implementation Details. For reproducibility of our main results we use the model Qwen2.5-VL-7B (Bai et al., 2025) trained on the **ReFLAIR-cold** dataset comprising 50K samples with responses formatted as <think><re-think><answer>. The Reflection Quality Scorer (RQS) is initialized from Qwen2.5-VL-3B. We employ a hybrid reward scheme based on GRPO for the reinforcement learning phase, which is built on the supervised fine-tuned model. Further implementation details are provided in Appendix B.

4.2 Main Results on Math Domain

Table 1 shows that *ReFLAIR* consistently demonstrates strong performance across all five benchmarks. On MathVista, MathVerse, and MM-Math—three challenging multimodal mathematical datasets—*ReFLAIR* reaches or surpasses the performance of both larger and similarly sized models. Compared with *ReFLAIR-SFT*, the reinforcement learning variant improves by 8.1 percentage points on MM-Math and 3.1 points on MathVerse, underscoring the importance of reward-guided reflection. On GSM8K and MATH500, *ReFLAIR* also outperforms most open-source baselines, including Qwen2.5-VL-7B-Instruct and Vision-R1-7B, and matches or exceeds the performance of models such as Satori-Qwen-7B that are also specifically designed for mathematics. These results confirm that our structured reflection method substantially enhances reasoning performance on both multimodal and symbolic mathematical tasks without increasing model size. Compared with proprietary models such as GPT-4o and OpenAI O1, *ReFLAIR*, despite being an open-source mid-scale model, achieves comparable or even superior performance, suggesting that structured reflection is a promising alternative to scale-intensive approaches.

4.3 Component Effects Across Tasks

<re-think> prompting. In the ablation setting, the model is constrained to predict directly with the <answer> token, effectively bypassing the reflective step. This shortcut removes the opportunity for the model to identify and revise intermediate reasoning errors. The comparison in Table 2 shows that eliminating the explicit reflection prompt consistently reduces performance, underscoring that structured prompting is indispensable for leveraging prior reasoning traces to improve final solutions.

Supervised cold-start training. Eliminating the initial supervised stage leads to a marked degradation on complex reasoning tasks. Such tasks often require multi-step symbolic manipulation or spatial inference, where naive trial-and-error provides little benefit. Training on high-quality reflection data establishes a prior for self-correction, enabling the model to perform meaningful adjustments during reinforcement learning rather than relying on random perturbations. The relatively smaller effect observed on GSM8K suggests that shorter and more structured problems can be effectively learned with reward-based optimization alone.

Reflection Quality Scorer (RQS). Substituting RQS with simple rule-based rewards yields consistent and substantial performance declines across benchmarks. This indicates that RQS provides a more precise alignment signal between reflection quality and model updates, particularly in datasets that contain ambiguous or open-ended reasoning errors. Although the gap is narrower on datasets such as GSM8K and MathVista—where carefully designed rules approximate reflection quality reasonably well—the overall results highlight that RQS enhances sensitivity to subtle yet meaningful improvements in reflection, especially those with semantic or logical significance.

Ablation analysis further reveals that the three components of ReFLAIR reinforce each other in a complementary manner rather than functioning in isolation. The <re-think> token provides the structural signal for reflection, but without supervised training on curated reflection data, the revisions remain shallow and often ineffective. Cold-start training establishes a prior for meaningful self-correction, which in turn enables RQS to serve as a precise reward signal rather than a noisy proxy. Likewise, RQS amplifies the benefits of reflection prompting by rewarding improvements that are semantically and logically substantive. Taken together, these components form a mutually reinforcing system. As shown in the final row of Table 2, the full ReFLAIR model achieves the best overall performance, with the largest margins observed on tasks requiring multi-hop or symbolic reasoning.

Model	MathVista	MathVerse	MM-Math	GSM8K	MATH500
Proprietary Models					
GPT-4V	49.9	39.4	23.1	-	-
GPT-4o	63.8	37.6	31.8	-	60.3
OpenAI O1	73.9	-	-	-	85.5
Open-Source Models					
Qwen2.5-VL-7B-Instruct	68.1	46.7	34.1	91.6	75.5
Qwen-2.5-Math-7B-Instruct	-	-	-	<u>95.2</u>	83.6
Qwen2.5-VL-72B	<u>73.5</u>	51.3	<u>45.6</u>	-	-
QwQ-32B-Preview	-	-	-	**95.5**	**90.6**
Vision-R1-7B	<u>73.5</u>	52.4	40.2	-	-
Satori-Qwen-7B	-	-	-	93.2	85.6
Ours					
ReFLAIR-SFT	73.1	<u>55.2</u>	38.2	93.1	82.8
ReFLAIR	**73.6**	**58.3**	**46.3**	93.9	<u>86.8</u>

Table 1: **Evaluation results on five mathematical reasoning benchmarks**: MathVista, MathVerse, MM-Math, GSM8K, and MATH500. We compare performance across proprietary models, large-scale open-source models, 7B-scale open-source models, and our ReFLAIR variants. **ReFLAIR** achieves consistently strong and often state-of-the-art performance, with clear advantages in multi-step mathematical reasoning.

Configuration	MathVista	MathVerse	MM-Math	GSM8K	MATH500
w/o <re-think>	71.3	53.6	40.5	92.3	81.3
w/o cold-start SFT	74.1	52.9	37.2	92.8	79.9
w/o RQS (rule only)	70.5	54.2	36.0	92.0	78.3
Full ReFLAIR	**73.6**	**58.3**	**46.3**	**93.4**	**84.2**

Table 2: Ablation study on the three key components of ReFLAIR across five benchmarks. Removing reflection prompting, cold-start supervised training, or RQS supervision all lead to notable performance degradation, especially on complex multimodal and symbolic reasoning tasks.

4.4 Reflection Quality Scorer (RQS) Analysis

4.4.1 Learning Dynamics of RQS Rewards

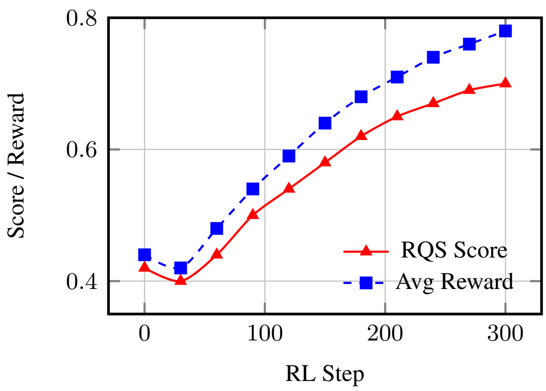

Figure 4: Training dynamics with RQS supervision. Both RQS scores and average rewards show a brief initial decline followed by consistent improvement, reflecting the exploration and optimization stages of reinforcement learning.

During reinforcement learning, both RQS scores and average rewards initially dip before steadily rising as training progresses. This pattern reflects the exploration phase of optimization, followed by convergence to higher-quality behaviors. The average reward remains consistently above the RQS score, and the gap widens over time, suggesting that improvements in reflection quality translate into progressively stronger task-level outcomes. This trajectory highlights the role of RQS in fostering reflections that are not only structurally coherent but also substantively beneficial for downstream reasoning.

4.4.2 Effect of Model Scale and Training Duration

Analysis of different RQS configurations shows that a smaller model trained for longer yields more reliable improvements than a larger model trained for the same or greater duration. The 3B model benefits from extended training, while the 7B variant shows signs of diminishing returns and even degradation under longer training. These findings suggest that reward quality depends more on careful calibration of training than on raw model capacity, with smaller models providing better generalization

RQS Configuration	MathVista	MathVerse	MM-Math	GSM8K	MATH500
3B, 1 epoch	73.1	57.9	43.2	92.8	79.4
3B, 2 epochs	**73.6**	**58.3**	**46.3**	**93.9**	**86.8**
7B, 1 epoch	73.4	57.2	45.7	92.3	81.9
7B, 2 epochs	72.1	56.0	43.9	92.0	80.6

Table 3: Downstream performance with different RQS reward models. The smaller 3B RQS trained for two epochs achieves the best overall performance, indicating that extended training is more influential than model size in producing a reliable reward signal.

when sufficiently optimized.

4.4.3 Validation of RQS Judgments

To assess the reliability of RQS, we constructed a set of reflective reasoning trajectories and obtained ground truth scores by applying the rule-based criteria described earlier. Three scoring sources were then compared against this ground truth: DeepSeek-R1, our learned RQS, and GPT-4o. Results are summarized in Table 4.

The agreement scores demonstrate that both DeepSeek-R1 and RQS deliver consistent evaluations of reflection quality, with RQS exhibiting slightly stronger alignment with the ground truth. This outcome shows that RQS does not merely replicate biases from its training source but successfully generalizes the reflection evaluation criteria, thereby providing a stable and trustworthy signal for reinforcement learning.

4.5 Test-Time Reflection Scaling

4.5.1 Fixed Reflection Depth

We first evaluate the effect of manually increasing the number of reflection steps at test time using the final trained *ReFLAIR* model. Experiments were conducted with one, two, and three reflection steps on MathVista and MM-Math. Results indicate that additional steps yield only marginal gains on MM-Math and no improvement on MathVista. Notably, performance deteriorates when three steps are enforced, primarily due to excessive verbosity and the accumulation of reasoning errors. Case analysis further reveals hallucination-like behaviors, where successive reflections compound spurious reasoning paths and lead to incorrect final answers. These findings suggest that unregulated overthinking can be detrimental, underscoring the need for mechanisms that balance depth and reliability in reflective inference.

4.5.2 Adaptive Reflection Scheduling

Since the optimal number of reflections depends strongly on task complexity, we further investigate an adaptive scheduling mechanism. A lightweight *Thought Quality Scorer* (TQS) is trained to assess the quality of reasoning fragments generated before the </think> token. Based on this score, the model either proceeds to an answer or initiates another reflection. This approach prevents unnecessary repetition on simple problems while allowing deeper reasoning when needed.

The adaptive strategy achieves measurable efficiency gains without compromising accuracy. On MathVista, it yields a modest but consistent performance increase, while average inference time is reduced by nearly a quarter. These results demonstrate that dynamically regulating reflection depth provides a principled balance between accuracy and computational cost.

5 Discussion

5.1 Discussion and Future Directions

Our results show that explicitly structured reflection improves both the accuracy and interpretability of large language model (LLM) reasoning. The <re-think> stage enables the model to revise or extend its initial reasoning, which is particularly valuable in multimodal settings where errors often arise from visual or symbolic misalignment. The reward design aligns cognitive plausibility with task-oriented metrics such as correctness and difficulty, offering a principled way to supervise beyond final answers.

ReFLAIR's modularity also supports integration with diverse prompting strategies and architectures. While we focus on mathematical and chart-based reasoning, the framework generalizes to other domains such as instruction following and critique generation, suggesting broader potential for reflection-aware training.

We further investigated whether reflection frequency can be learned end-to-end through reinforcement learning. After equipping the model with multi-step reflective reasoning via supervised fine-tuning, RL was used to associate task difficulty

Scorer	Ground Truth (%)	GPT-4o Agreement (%)
DeepSeek-R1	93.25 / 98.50	95.00 / 97.25
RQS	95.75 / 98.75	95.25 / 99.00

Table 4: Agreement between RQS, DeepSeek-R1, and GPT-4o on reflection quality ratings. Results indicate that RQS provides robust and generalizable judgments closely aligned with the rule-based ground truth.

Reflection Steps	MathVista	MM-Math
One step	73.6	46.3
Two steps	73.6	47.2
Three steps	71.9	45.3

Table 5: Performance impact of enforcing fixed numbers of reflection steps. Excessive iterations introduce verbosity and reasoning drift, leading to degraded results.

Method	MathVista	Avg. Inference Time
Fixed (1 step)	73.6	1.0×
Fixed (2 steps)	73.6	1.8×
Adaptive (TQS)	74.5	0.77×

Table 6: Comparison of fixed and adaptive reflection strategies. The TQS-based mechanism achieves higher efficiency while maintaining or improving accuracy.

with reflection depth. Across five benchmarks, however, the improvements over fixed single-step inference were marginal. This suggests that while RL can in principle capture scheduling behavior, more refined formulations are needed. In contrast, our adaptive scheduling approach based on lightweight thought-quality estimation already achieves measurable gains in both accuracy and efficiency.

These findings point toward several directions: richer forms of meta-reflection spanning multiple <re-think> cycles, adaptive control informed by task complexity or confidence, and alignment of reflective reasoning with human pedagogical preferences. Advancing along these axes may yield systems that are not only more accurate but also more interpretable and aligned with human reasoning practices.

5.2 Generalization to Large Language Models (LLMs)

To examine the applicability of our approach beyond multimodal reasoning, we further evaluate ReFLAIR on a custom-built textual math question answering dataset using purely language-based LLMs. The results, summarized in Table 7, show that ReFLAIR provides substantial improvements, especially for smaller models. In particular, Qwen2.5-7B-Instruct achieves an absolute gain of +35.3 points under ReFLAIR compared to its baseline, a margin even larger than the improvements observed with multimodal models. This result highlights both the difficulty of the dataset and the effectiveness of reflection-aware training in enhancing the reasoning capacity of compact models. We will release this dataset to encourage further investigation and reproducibility.

Model	Accuracy (%)
o1-preview	60.2
Qwen2.5-7B-Instruct	21.0
Qwen2.5-72B-Instruct	52.9
ReFLAIR-SFT	48.9
ReFLAIR	**56.3**

Table 7: Performance comparison on a custom-built textual math QA dataset using purely textual LLMs. ReFLAIR provides substantial improvements, particularly for smaller-scale models.

6 Conclusion

We have introduced **ReFLAIR**, a framework that elevates reflective reasoning to a central component of multimodal language model training. ReFLAIR integrates three ingredients: a large-scale reflection-annotated dataset for cold-start supervision, a trainable Reflection Quality Scorer that quantifies the utility of rethinking steps, and a hybrid reinforcement learning scheme that balances correctness, structural fidelity, and reflection utility. Experiments on diverse mathematical reasoning benchmarks demonstrate consistent gains in both accuracy and interpretability, supported by ablation studies, transfer experiments to textual models, and analyses of test-time scaling. These results highlight that structured reflection, when aligned with explicit utility signals, provides a scalable path toward reasoning systems that critique, revise, and improve their own cognition.

Limitations

Despite its advantages, ReFLAIR has several limitations. First, incorporating reflection introduces additional inference cost, which is mitigated but not eliminated by the proposed adaptive scheduling mechanism. More principled approaches to dynamically deciding when reflection is necessary remain to be explored. Second, the Reflection Quality Scorer is trained on proxy annotations derived from model trajectories. Although validated against human judgments, its alignment with human preference and pedagogical value is still limited. Third, the framework has been primarily trained and evaluated on math-focused multimodal data. Its extension to other domains such as dialogue, programming, or long-context reasoning tasks is yet to be demonstrated. Finally, reflection itself is shaped by cultural, ethical, and stylistic norms, and optimizing behavior in line with these diverse considerations remains an open challenge for deployment in real-world applications.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ming Chen, Hao Xu, and Faye Li. 2025. Lr^{2}bench: A benchmark for long-chain reflective reasoning. In *Findings of the Association for Computational Linguistics (ACL)*.

Karl Cobbe, Vova Kosaraju, Miljan Bavarian, Curtis Ellis, Caglar Gulcehre, and John Schulman. 2021. Training verifiers to solve math word problems. In *ICLR*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Yuxuan Fu, Qianxi Zhang, Biqing Wei, Xiang Sun, Hui Huang, Feng Zhuang, Zhiwei Huang, Shiliang Huang, and Peng Zhao. 2023. Mme: A comprehensive evaluation benchmark for multimodal llms. *arXiv preprint arXiv:2306.13394*.

Shuangfei Guo, Tianjian Li, Jing Sun, Ruiqi Zhang, and Momin Yang. 2025. Advancing language model reasoning through reinforcement learning and inference scaling. *arXiv preprint arXiv:2501.11651*.

Himanshu Gupta, Shreyas Verma, Ujjwala Anantheswaran, Kevin Scaria, Mihir Parmar, Swaroop Mishra, and Chitta Baral. 2024. Polymath: A challenging multi-modal mathematical reasoning benchmark. *arXiv preprint arXiv:2410.14702*.

Dan Hendrycks, Collin Burns, Sebastian Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models.

Yue Huang, Hang Li, Yukun Chen, Xin Chen, Yi Ma, Fajie Song, Yu Zheng, Duyu Li, and Maosong Sun. 2024a. Mathverse: A unified multimodal math benchmark for real-world problem solving. *arXiv preprint arXiv:2404.02615*.

Yue Huang, Hang Li, et al. 2024b. Mathverse: A unified multimodal math benchmark for real-world problem solving. *arXiv preprint arXiv:2404.02615*.

Jia Li, Edward Beeching, Lewis Tunstall, Benjamin Lipkin, Roman Soletskyi, Shengyi C. Huang, Kashif Rasul, Alexander Jiang, Ziju Shen, Zihan Qin, Le Zhou, Yuhui Fleureau, Guillaume Lample, and Sam Polu. 2024a. Numinamath: The largest public dataset in ai4maths with competition-level math problems and chain-of-thought solutions. NuminaMath-CoT dataset, approximately 860K problem-solution pairs; arXiv preprint or dataset URL.

Yiming Li, Xia Chen, Liang Zhou, and Zheng Wang. 2024b. Hindsight: Selective reflection for robust reasoning. *arXiv preprint arXiv:2406.12050*.

Yiming Li, Wei He, Boyang Hong, Shihan Do, and Wenyu Zhan. 2025. Two heads are better than one: Decoupling critique and refinement for enhanced reasoning. *arXiv preprint arXiv:2501.23456*.

Haotian Liu, Xiaohong Li, et al. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *arXiv preprint arXiv:2404.09860*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning in visual contexts. In *ICLR*.

Arjun Madaan, Aakanksha Srinivasan, Ammar Wang, and Uri Hasson. 2023. Self-refine: Iterative refinement using self-feedback for reasoning. In *Findings of ACL*.

Mahmoud Masry, Enamul Hoque, and Giuseppe Carenini. 2021. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *EMNLP*.

OpenAI. 2024. Introducing openai o1. Accessed: 2025-05-20.

Leonardo Ranaldi and Andr'e Freitas. 2024. Self-refine instruction-tuning for aligning reasoning in language models. In *EMNLP*.

Maohao Shen, Guangtao Zeng, Zhenting Qi, Zhang-Wei Hong, Zhenfang Chen, Wei Lu, Gregory Wornell, Subhro Das, David Cox, and Chuang Gan. 2025. Satori: Reinforcement learning with chain-of-action-thought enhances llm reasoning via autoregressive search.

Thomas Shinn, Cedric Li, Fanie Dalvi, Akari Asai, Joseph Eichstaedt, Barry Haddow, Edgar Lambert, Mikhail Mishra, Tanner Spero, William Wang, and Gregory Wakefield. 2023. Reflexion: Language agents with verbal self-reflection improve tool use. *arXiv preprint arXiv:2307.02415*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize with human feedback. In *NeurIPS*.

Xiaoqing Sun, Weilin Zhang, Xiaoyan Jin, Yiming Tsai, Devi Parikh, Kai Wang, Hendrik Becher, Fei-Fei Li, Song-Chun Yu, Yan Tian, and colleagues. 2024. Mm-star: A multimodal benchmark for star-level evaluation. *arXiv preprint arXiv:2402.00001*.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. 2024. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ke Wang, Junting Pan, Linda Wei, Aojun Zhou, Weikang Shi, Zimu Lu, Han Xiao, Yunqiao Yang, Houxing Ren, Mingjie Zhan, and Hongsheng Li. 2025. Mathcoder-VL: Bridging vision and code for enhanced multimodal mathematical reasoning. In *The 63rd Annual Meeting of the Association for Computational Linguistics*.

Yujie Wang, Xiaonan Zhou, Dashi Huang, Shaobo Wu, Yixin Zheng, and Xilin Lu. 2024b. Mm-math: A benchmark for multimodal math problem solving. In *EMNLP*.

Yujie Wang, Xiaonan Zhou, et al. 2024c. Mm-math: A benchmark for multimodal math problem solving. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS) 35*.

Zhenyu Yan, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. 2024. Mirror: Diverse reflective reasoning without ground-truth access. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

Victor Zelikman, Omer Levy, Hannaneh Hajishirzi, Luke Zettlemoyer, and Ian Tenney. 2022. Star: Self-taught reasoner improves logical reasoning in language models. *arXiv preprint arXiv:2203.14465*.

Yao Zhang, Michael Tang, Zhe Gan, Xiang Wang, James Glass, Wei Xu, and Louis Rosenberg. 2022. Least-to-most prompting enables complex reasoning in large language models. In *NeurIPS*.

A Illustrative Reflective Trajectories

We present two examples that showcase different roles of structured reflection. The first example demonstrates how a brief self-review corrects a shallow but common error. The second example highlights a distinct pattern: the model uses the initial <think> stage to outline a multi-step plan without performing detailed computations, and the subsequent <re-think> stage executes that plan to completion. No external images are required for understanding these examples; the captions emphasize how reflection reduces superficial mistakes and converts planning into reliable execution.

This example illustrates the corrective role of reflection. The initial step applies a superficial rule that is easy to produce but incorrect in context. The reflection stage explicitly recalls the appropriate quantity to compute, derives the total time and distance, and produces the correct result. The trajectory is short and focused: reflection targets the precise misconception rather than expanding verbosity.

The second example demonstrates a complementary use of reflection. During <think>, the model constructs a structured plan that enumerates the counting stages and potential pitfalls. By deferring arithmetic to <re-think>, the model reduces the likelihood of early calculation errors and explicitly validates modeling assumptions. The executed trajectory confirms that the correct combinatorial decomposition is $\binom{3}{2} \times 5 \times 4 = 60$, and it avoids double-counting because the collaborative assignment is treated as a set assigned to a single, labeled project while the remaining student is assigned to a distinct, labeled project.

Input Question:
A train travels from City A to City B at an average speed of 60 km/h. It returns from City B to City A on the same route at an average speed of 40 km/h. What is the average speed of the entire round trip?
Model Reasoning:
<think> To find the average speed of the entire round trip, I can simply average the two speeds: (60 + 40) / 2 = 50 km/h. </think>
<re-think> The arithmetic mean of the two speeds is not the correct measure for average speed over a round trip. Average speed should be total distance divided by total time.
Let the one-way distance be d km. The time for the first leg is $d/60$ hours, and for the return leg it is $d/40$ hours. The total distance is $2d$ and the total time is $d/60 + d/40 = 5d/120 = d/24$ hours. Therefore the average speed is

$$\frac{2d}{d/24} = 48 \text{ km/h}.$$

</re-think>
<answer> The average speed for the round trip is **48 km/h**. </answer>

Table 8: A concise corrective reflection. The initial <think> adopts a common but shallow heuristic, while the <re-think> stage replaces the heuristic with a distance–time derivation that yields the correct answer.

B Implementation Details

In this appendix we provide full specifications of training configurations and dataset splits that underpin the results reported in Section 4.

The **ReFLAIR-cold** dataset comprises 50,000 examples randomly sampled (without replacement) from the union of MathVision, PolyMath, OpenMathInstruct-2, and NuminaMath-CoT, after quality filtering to remove examples with malformed reasoning or invalid answers. For **ReFLAIR-RQS** and **ReFLAIR-RL**, the same sources are used; their splits are 15,000 examples in the RQS training set (for annotated reflections) and approximately 15,000 examples for the RL phase.

All fine-tuning is performed with the AdamW optimizer, learning rate set to 5×10^{-5}, weight decay 0.01, batch size 64. Qwen2.5-VL-7B is trained for one epoch over ReFLAIR-cold, using 8 NVIDIA A100 GPUs, which amounts to roughly 22 hours of wall-clock time. Qwen2.5-VL-3B is trained with identical hyperparameters. The Reflection Quality Scorer (RQS) is initialized from Qwen2.5-VL-3B and fine-tuned for two epochs on 15,000 annotated reflections, using 2 A100 GPUs, taking approximately 12 hours. For reinforcement learning (ReFLAIR-RL), we adopt a GRPO-based algorithm with a hybrid reward composed of correctness of the final answer and a quality metric of reflection steps; this phase builds on the SFT model and is run for approximately 36 hours on 8 A100 GPUs. Gradient clipping is applied at norm 1.0; mixed precision (FP16) is used throughout. All training seeds are fixed for reproducibility.

Input Question:
Five distinct projects are available and three students are to be assigned so that one project requires a pair of students working together and one other project accepts exactly one student. In how many distinct ways can the students be assigned under these constraints?
Model Reasoning:
<think> I will outline a plan before computing. First, choose which two students will form the collaborating pair. Next, select which project will host the collaborating pair. Then choose the project that will receive the remaining student. I will not carry out the arithmetic here; instead I will follow this plan and perform each counting step carefully in the revision stage. </think>
<re-think> Execute the planned steps precisely. Choose the pair of students in $\binom{3}{2} = 3$ ways. Choose the project for the collaborating pair in 5 ways. From the remaining projects, choose the project for the single student in 4 ways. The collaborative pair is assigned as an unordered set to its project, so no additional internal permutation should be counted. Multiplying these factors yields

$$3 \times 5 \times 4 = 60.$$

Thus there are **60** distinct valid assignments under the stated constraints. </re-think>
<answer> There are **60** distinct valid assignments. </answer>

Table 9: A multi-step planning and execution example. The <think> stage states a clear plan without numerical work; the <re-think> stage follows that plan, performs the required combinatorial computations, and verifies assumptions about ordering and distinguishability.