

Knowing What to Solve Before How: Preplan Empowered LLM Mathematical Reasoning

Shaojie Wang¹ Liang Zhang^{1*}

{shaojiewang, liangzhang}@hkustgz.edu.cn

¹ Hong Kong University of Science and Technology (Guangzhou)

Abstract

Current plan-based reasoning methods improve large language models (LLMs) by inserting a planning stage before execution, giving rise to the **question** → **plan** → **cot** paradigm. While effective, a closer examination reveals an inherent paradigm-level gap: both the planning and its execution stages decide *how to solve* a problem, while the prior question of *what to solve*; recognizing the problem type, the applicable tools, and the foreseeable pitfalls; remains entirely implicit. To bridge this gap, we propose **PPC** (Preplan-Plan-CoT), a framework that introduces an explicit problem-understanding stage, the *preplan*, yielding a new **question** → **preplan** → **plan** → **cot** paradigm. Realizing this paradigm requires safeguarding the conceptual integrity of preplan at both ends. Specifically, we design a three-stage synthesis pipeline with a *spoiler-score detector* that filters out leakage and spoiler failures to build clean preplan supervision, and a composite GRPO reward enforces that the generated plan genuinely follows from the preplan. Experiments across four backbones and five mathematical reasoning benchmarks show that PPC achieves the best results on 39 of 40 metrics, improving *maj@16* and *pass@16* by +2.23 and +3.06 over the strongest baseline without introducing additional inference token overhead.

1 Introduction

Large language models (LLMs) have made remarkable progress on mathematical reasoning tasks, driven by the Chain-of-Thought (CoT) paradigm (Wei et al., 2022; Kojima et al., 2022). Under this **question** → **cot** paradigm, the model produces, in a single forward pass, a long sequence that interleaves intermediate reasoning steps with the final answer. Combined with supervised fine-tuning (SFT) on CoT trajectories (Yue et al., 2024) and reinforcement learning (RL) from verifiable

rewards (Guo et al., 2025), this paradigm has achieved strong results across a wide range of mathematical benchmarks (Ahn et al., 2024).

As problems grow more complex, however, single-pass CoT reveals a structural limitation: operating as an autoregressive, token-level process, it confines reasoning to *step-level* execution and lacks higher-level organization, yielding redundant or incoherent trajectories. Recent work addresses this issue by inserting an intermediate planning stage, giving rise to the **question** → **plan** → **cot** paradigm: *Plan-Tuning* (Parmar et al., 2025) distills planning trajectories from large models and fine-tunes smaller LLMs on (question, plan, solution) triples, while *PTA-GRPO* (Dou et al., 2025) introduces plan-level guidance rewards during reinforcement learning. Their shared insight is that an explicit *plan-level* layer above *step-level* execution supplies the global organization that single-pass CoT lacks.

However, a closer examination of the **question** → **plan** → **cot** paradigm reveals that its two stages, planning and execution, differ in granularity but share the same underlying role: both decide *how to solve* the problem—the plan organizes the solution, and the execution carries it out. What neither stage addresses is the prior question of *what to solve*: such as recognizing what type of problem it is, which tools or concepts may apply, which boundary conditions matter, and which pitfalls to anticipate. For example, given a problem $12x^2 - xy - 6y^2 = 0$, recognizing it as a homogeneous quadratic immediately signals the model to prioritize factorization over the more cumbersome discriminant method, while also indicating that each linear factor produces a family of solutions. Consequently, the subsequent plan can follow these hints and count solutions correctly. Existing methods implicitly assume that such understanding emerges for free during plan generation, leaving a **paradigm-level gap**: *how to solve* is modeled explicitly, whereas

* Corresponding author.

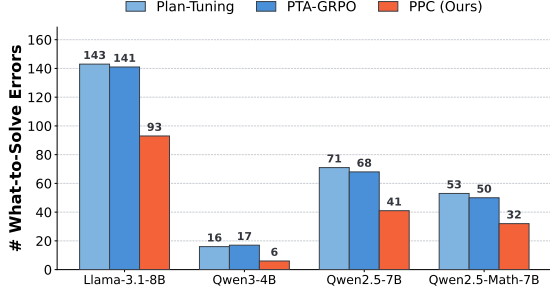


Figure 1: Number of *what-to-solve* errors on MATH-500 across four backbones. Each wrong answer (under greedy decoding) is attributed to its root cause by an LLM judge, here we use DeepSeek-V4.

what to solve remains implicit. We refer to this missing stage as the *preplan*: an explicit problem-understanding stage that precedes planning.

Figure 1 illustrates the cost of this paradigm-level gap. We diagnose the errors made by two representative **question** → **plan** → **cot** methods on MATH-500 (Lightman et al., 2023), using an LLM judge to attribute each wrong answer to its root cause (see Appendix A for more details). Across four backbones, a large fraction of their errors stem not from calculation, but from a failure to understand *what to solve*. This exposes the cost of leaving problem understanding implicit and motivates introducing it as an explicit stage prior to planning.

While conceptually intuitive, bridging this gap by directly introducing a problem-understanding stage faces two concrete challenges. **Challenge 1 (data construction)**. No existing reasoning dataset provides supervision for such a preplan stage, and directly prompting a strong LLM to generate it tends to fail in two ways. The output either rehearses the forthcoming plan (*leakage failure*), or quietly carries out intermediate computation under the guise of analysis (*spoiler failure*). Both collapse the stage back into *how to solve*, defeating its purpose. **Challenge 2 (training-time adherence)**. Even given clean supervision, a model trained with standard outcome rewards may learn to emit a superficially well-formed analysis while ignoring it during plan generation, capturing the mere form of problem understanding without its practical utility. A successful framework must therefore enforce that the downstream plan genuinely follows from the preceding preplan.

To address these challenges, we propose **PPC (Preplan-Plan-CoT)**, a framework that explicitly introduces a problem-understanding stage, the *preplan*, into both data construction and reinforcement learning, giving rise to the new **question** → **pre-**



Figure 2: An example comparing the two paradigms on the same problem. *Top*: a **question** → **plan** → **cot** solver commits to the discriminant route and undercounts (84). *Bottom*: a preceding *preplan* recognizes the homogeneous form, chooses factoring over the fragile discriminant route, and notes that each factor is a whole family of lattice points; the plan inherits this and the count is correct (117). Full trace in Appendix B.

plan → **plan** → **cot** paradigm. A preplan is a comprehensive analysis about the problem that identifies the problem type, surveys relevant tools and concepts, flags the boundary conditions that matter, and anticipates pitfalls—it settles *what to solve* and leaves *how to solve* to the subsequent planning and execution stages. To address Challenge 1, we design a three-stage synthesis pipeline guarded by a *spoiler-score detector* that filters out leakage and spoiler failures, yielding a clean reasoning dataset with explicit preplan supervision. For Challenge 2, we extend GRPO with a composite reward that jointly optimizes *final-answer correctness* and *plan-preplan alignment*, while enforcing structural compliance across the three stages and penalizing preplan style degradation through the same spoiler score.

Figure 2 makes this new paradigm concrete: on the same problem, a **question** → **plan** → **cot** solver commits to a fragile route without analyzing what the problem requires, whereas an explicit preplan

first recognizes the structure and selects a reliable route, guiding the subsequent plan toward the correct solution. This mirrors the aggregate trend in Figure 1. Our main contributions are as follows:

- Conceptually: We identify and formalize a **paradigm-level gap** in existing plan-based reasoning: current methods model *how to solve* explicitly but leave *what to solve* implicit. To bridge it, we propose the **question** $\rightarrow$ **preplan** $\rightarrow$ **plan** $\rightarrow$ **cot** paradigm for the first time.
- Methodologically: We construct the first reasoning dataset with explicit *preplan* supervision, protected by a *spoiler-score detector* against the two characteristic generation failures, leakage and spoiler. We then design a composite GRPO reward whose *plan-preplan alignment* term explicitly enforces that the generated plan follows from the preceding preplan, preventing the model from generating superficially well-formed but otherwise ignored preplans.
- Empirically: Across four backbones and five mathematical reasoning benchmarks, PPC achieves the best results on 39 of 40 metrics, improving maj@16 and pass@16 by +2.23 and +3.06 over the strongest baseline without introducing additional inference token overhead.

2 Related Work

2.1 LLM Reasoning with Chain-of-Thought

Chain-of-thought (CoT) prompting (Wei et al., 2022; Kojima et al., 2022) established that eliciting intermediate reasoning steps before the final answer substantially improves LLM performance on complex reasoning tasks, and has since become the dominant paradigm for mathematical reasoning. To strengthen this capability beyond in-context prompting, two complementary post-training directions have been actively explored. The first relies on supervised fine-tuning over distilled CoT trajectories (Guha et al., 2025; Ye et al., 2025; Moshkov et al., 2025), where strong teacher models generate large pools of step-by-step solutions that are used to fine-tune smaller students. The second, popularized by DeepSeek-R1 (Guo et al., 2025), and subsequently extended by a growing line of work (Shao et al., 2024; Chen et al., 2026; He et al., 2025a), applies reinforcement learning with verifiable rewards (RLVR), in which models receive

direct correctness feedback from deterministic verifiers and learn to produce long CoT trajectories with exploration. Despite these advances, recent analyses show that CoT trajectories generated by LLMs remain *step-level*: the model derives each step locally as it decodes, without an explicit mechanism to commit to a global solution structure beforehand (Dou et al., 2025; Wang et al., 2023). This limitation motivates a complementary line of work that introduces higher-level structure into the reasoning process.

2.2 Plan-Based Reasoning

A complementary direction to scaling step-level CoT is to introduce *higher-level structure* into the reasoning process. Early prompting-based work asks the model to plan before solving (Wang et al., 2023) or decomposes problems into simpler sub-questions (Zhou et al., 2022), with later extensions exploring hierarchical sub-goals (Gui et al., 2025; Huang et al., 2026) and tree-structured exploration (Yao et al., 2023). They share the premise that organization above individual steps improves coherence on multi-stage problems. Recent work integrates planning into post-training. Plan-Tuning (Parmar et al., 2025) distills planning trajectories from large reasoning models and supervises smaller LLMs on (question, plan, solution) triples, while PTA-GRPO (Dou et al., 2025) extends this to RL with a plan-aware reward that jointly evaluates plan quality and final answer correctness, elevating planning from an inference-time trick to a training-time objective. Yet all of these methods treat the *plan* as the entry point of structured reasoning: both the planning and its execution answer *how to solve the problem*, while the prior question of *what* the problem is asking remains implicit. Our work targets this gap by introducing an explicit preplan stage upstream of planning, together with a training objective that enforces the downstream plan to genuinely follow from it.

3 Methodology

Our goal is to teach LLMs a new reasoning paradigm, **question** $\rightarrow$ **preplan** $\rightarrow$ **plan** $\rightarrow$ **cot**, in which a *preplan* that analyzes *what to solve* precedes the planning and execution that decide *how to solve*. The defining property of the preplan is that it analyzes *what to solve* rather than *how*: it surveys problem type, tools, constraints, and pitfalls without committing to a derivation. This conceptual boundary is fragile at both ends of training. During

data construction, a generator asked for a preplan tends to slip back into *how to solve*, through leakage or spoiler; during **reinforcement learning**, a policy may reproduce the *form* of a preplan while ignoring it when generating the following plan: the two failures of Challenge 1 and Challenge 2 (§1). PPC addresses both through a single organizing principle. Stage 1 (§3.2) supplies a *clean demonstration* of the paradigm by filtering out leakage and spoiler, and Stage 2 (§3.3) enforces *faithful use* through a composite reward. Crucially, the same rule-based signal $s(\cdot)$ patrols both ends: as a hard filter when building data (Eq. 4) and as a soft penalty during RL, ensuring that the same conceptual integrity constraint is enforced.

3.1 Preliminaries

Problem formulation. Given a question q , the policy π_θ generates a structured trajectory $y = (y_{pp}, y_p, y_e)$ enclosed by tags `<preplan>`, `<plan>`, `<execute>`, with the final answer $\hat{a}(y)$ parsed from a `\boxed{\cdot}` marker in y_e . Unlike the prior **question**→**plan**→**cot** paradigm, we additionally require y_{pp} to be a non-computational problem-understanding analysis upstream of the plan.

GRPO. We adopt Group Relative Policy Optimization (Shao et al., 2024) as our RL backbone. For each prompt q , GRPO samples a group $\{y^{(i)}\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot \mid q)$ and computes group-normalized advantages

$$\hat{A}_i = \frac{R(y^{(i)}) - \text{mean}(\{R(y^{(j)})\}_{j=1}^G)}{\text{std}(\{R(y^{(j)})\}_{j=1}^G)} \quad (1)$$

then optimizes the clipped policy objective with a KL regularizer to a reference policy π_{ref} :

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q, \{y^{(i)}\}} \left[\frac{1}{G} \sum_{i=1}^G \min(\rho_i \hat{A}_i, \text{clip}(\rho_i, 1 - \epsilon, 1 + \epsilon) \hat{A}_i) - \beta \mathbb{D}_{\text{KL}}[\pi_\theta \parallel \pi_{\text{ref}}] \right] \quad (2)$$

where $\rho_i = \pi_\theta(y^{(i)} \mid q) / \pi_{\theta_{\text{old}}}(y^{(i)} \mid q)$ and ϵ, β are hyperparameters.

3.2 Stage 1: Preplan-Supervised Data Construction

The difficulty for dataset construction is not generation but *conceptual integrity*. Specifically, a *leakage* preplan typically rehearses the plan’s step sequence; while a *spoiler* preplan tends to substitute concrete computation for structural description. Both collapse the stage back into *how to solve*. Our

pipeline counters the first *by construction* and the second *by filtering*.

Three-stage synthesis. Each trajectory is generated by conditioning every stage only on its predecessors:

$$\begin{aligned} y_{pp} &\sim \pi_{pp}(\cdot \mid q), & y_p &\sim \pi_p(\cdot \mid q, y_{pp}), \\ y_e &\sim \pi_e(\cdot \mid q, y_p) \end{aligned} \quad (3)$$

The strict left-to-right, stage-wise generation process allows finer-grained control over different components of the reasoning trajectory. Leakage is suppressed *by construction* at the prompt level: π_{pp} is prompted to forbid derivations, intermediate values, step-by-step procedures, and any reference to a solution path, so that the supervision targets *what to solve* rather than *how*. Implementation details are given in Appendix C.

Spoiler-score filtering. Left-to-right ordering does not prevent *spoiler*, since a generator can compute within a single stage. We therefore retain a trajectory only if its preplan is pure and its answer is correct:

$$s(y_{pp}) \leq \tau_s \quad \text{and} \quad \hat{a}(y) \equiv a^* \quad (4)$$

where $s(\cdot) \in \{0, \dots, 6\}$ is a rule-based *spoiler score* that aggregates several derivation- and answer-revealing signals (defined in Appendix D). The check is deliberately decoupled from correctness: a trajectory whose answer is correct is *still discarded* if its preplan is impure. For instance, a preplan that already asserts computed classifications and rehearses the exact series-expansion procedure scores high and is filtered out despite reaching the correct answer (full example in Appendix B). This separation is the point: the filter targets preplan purity, not answer correctness. The retained corpus $\mathcal{D}_{\text{SFT}}$ initializes the policy by token-level fine-tuning,

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(q, y) \sim \mathcal{D}_{\text{SFT}}} \sum_t \log \pi_\theta(y_t \mid q, y_{<t}) \quad (5)$$

yielding the model π_{θ_0} that initializes the RL stage.

Data quality audit. To verify that the dataset indeed meets the intended design, we additionally conduct an independent LLM-as-judge audit: over 500 sampled trajectories across three seeds, 97.6–98.4% pass a four-criterion check covering preplan purity, plan–preplan faithfulness, execution correctness, and stage-wise coherence, confirming the pipeline reliably yields clean preplan supervision.

3.3 Stage 2: Composite-Reward RL

Clean demonstrations teach the policy to produce the paradigm’s form, but imitation cannot guarantee the policy actually *uses* each stage: a fluent preplan may be silently ignored when the plan is generated (Challenge 2). We therefore use reinforcement learning to anchor every stage to an explicit reward, replacing the binary outcome reward of standard GRPO with a four-term composite:

$$R(y) = R_{\text{out}}(y) + \lambda_a R_{\text{adh}}(y) + \lambda_f R_{\text{fmt}}(y) - \lambda_s R_{\text{sty}}(y) \quad (6)$$

with weights $\lambda_a, \lambda_f, \lambda_s > 0$. Intuitively, R_{out} anchors *still solving the problem*, R_{adh} anchors *the preplan being inherited by the plan*, and $R_{\text{fmt}}, R_{\text{sty}}$ prevent the first two from being gamed. All scores come from *frozen* judges not updated during RL; exact mappings, weights, and judge prompts are in Appendix E, with ranges summarized in Table 1.

(1) Outcome reward with partial credit. A 0/1 reward gives no gradient on near-miss trajectories whose preplan and plan are nonetheless informative. Following the partial-credit view of rubric-based rewards (Gunjal et al., 2025), we score incorrect answers by an LLM-judged proximity level $J_{\text{prox}}(y)$ rating *solution-path quality* (not numerical closeness):

$$R_{\text{out}}(y) = \begin{cases} 1, & \hat{a}(y) \equiv a^*, \\ g(J_{\text{prox}}(y)), & \text{otherwise} \end{cases} \quad (7)$$

where g is a monotone map capped strictly below 1. The cap preserves a margin between near-miss and correct trajectories, thus the policy cannot inflate partial credit into the appearance of a correct answer. We further verify partial credit’s utility and reliability in Appendix E.

(2) Plan–preplan adherence reward. An LLM critic $J_{\text{adh}}(\cdot, \cdot)$ scores the *strategic alignment* between y_{pp} and y_p , normalized to $R_{\text{adh}}(y) \in [0, 1]$. The critic is instructed to judge alignment, *not* plan quality; thus, a strong plan that ignores the preplan still obtains lower score. This also prevents the preplan from being decoratively produced but functionally ignored: tying the plan to the preplan rather than letting it become an independent solution.

(3) Structural format guard. $R_{\text{fmt}}(y) \in \{0, 1\}$ checks that the three tags each appear exactly once and in order and that a `\boxed{\}` marker lies inside y_e , preventing the other rewards from being satisfied by malformed outputs.

Table 1: Reward components, their ranges, and roles. Correctness dominates; adherence nudges; format and style act as guards.

Term	Range	Role
R_{out}	$[0, 0.5]$	main signal (partial credit)
$\lambda_a R_{\text{adh}}$	$[0, \lambda_a]$	adherence nudge
$\lambda_f R_{\text{fmt}}$	$\{0, \lambda_f\}$	structural guard
$-\lambda_s R_{\text{sty}}$	$[-\lambda_s \cdot (6 - \tau_s), 0]$	anti-degradation

(4) Spoiler-style penalty. Reusing the spoiler score $s(\cdot)$ from Eq. (4), we apply a one-sided penalty $R_{\text{sty}}(y) = \max(0, s(y_{\text{pp}}) - \tau_s)$, activated only when the preplan drifts back toward derivation-heavy content during RL. Thus the conceptual integrity of preplan purified at construction time is held throughout training by the same signal.

Summary R_{out} keeps correctness as the dominant objective; R_{adh} ensures the preplan is *used* rather than *produced*; R_{fmt} and R_{sty} prevent these signals from being satisfied through ill-formed outputs or by collapsing the preplan into computation. Notably, without R_{sty} the model can satisfy adherence by writing a derivation-style preplan that trivially aligns with the downstream plan, defeating the purpose of our paradigm. We ablate each term in §4.2.2.

4 Experiments

4.1 Experimental Setup

Training data. We build our training corpus from DeepMath-103K (He et al., 2025b), a large-scale mathematical reasoning dataset with per-problem difficulty annotations. We retain problems spanning four difficulty levels from medium to competition-hard, and stratify-sample a non-overlapping SFT/RL split of problems. SFT trajectories are then synthesized and filtered as described in §3.2, yielding the final training set.

Models. To evaluate generality across model scales and families, we apply our framework to four base models: Qwen2.5-7B-Instruct (Team et al., 2024), Qwen2.5-Math-7B-Instruct (Yang et al., 2024), LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen3-4B-Instruct (Yang et al., 2025).

Benchmarks. We evaluate on five mathematical reasoning benchmarks spanning a wide difficulty range: AIME25, Minerva-Math (Lewkowycz et al., 2022), OlympiadBench (He et al., 2024), MATH-500 (Lightman et al., 2023) and GSM8K (Cobbe et al., 2021).

Baselines. We compare against the following methods, all trained on the same data splits for fairness: (i) **Base**: the untrained backbone with zero-shot CoT prompting; (ii) **Prompt-Only**: the same untrained backbone prompted to first analyze the problem itself: its type, relevant concepts, strategic direction, constraints, and pitfalls; and then form a plan before solving; this elicits PPC’s preplan-then-solve structure purely through instruction, isolating whether prompt engineering alone suffices and thereby testing the necessity of our training; (iii) **GRPO** (Shao et al., 2024): standard GRPO with a binary outcome reward, initialized from vanilla SFT; (iv) **Plan-Tuning** (Parmar et al., 2025): a post-training framework that distills planning trajectories from larger LLMs and fine-tunes the model via combined supervised and reinforcement-learning objectives; (v) **PTA-GRPO** (Dou et al., 2025): a two-stage framework that first performs SFT on distilled high-level guidance and then applies GRPO with a plan-aware reward.

Evaluation protocol. For each problem we sample $k = 16$ trajectories and report *maj@16* (self-consistency majority vote) as our primary metric, together with *pass@16* as an upper-bound coverage measure. All methods use the identical chat template and system prompt (“*Please reason step by step, and put your final answer within \boxed{\}*”). Final answers are extracted from the last `\boxed{\}` marker and scored by the same answer-equivalence verifier used in training (§3.3).

Implementation details. **Data synthesis:** The three-stage trajectories of §3.2 are generated with Qwen3-235B (Yang et al., 2025) as the preplan and plan generators (π_{pp} , π_p) and DeepSeek-R1 (Guo et al., 2025) as the execution model (π_e); the raw solution is then reorganized into a numbered execution by the same instruction-tuned model. The spoiler filter (Eq. 4) uses threshold $\tau_s = 2$ and retains preplans of length $\ell_{\min} = 150$ to $\ell_{\max} = 1500$ tokens.

Training: We initialize from each backbone via SFT on $\mathcal{D}_{\text{SFT}}$ for 3 epochs with learning rate $1e - 5$ and batch size 16, then run composite-reward GRPO for 500 steps. GRPO uses group size $G = 8$ and reward weights in Eq. (6) are $\lambda_a = 0.1$, $\lambda_f = 0.3$, and $\lambda_s = 0.1$. We verify in Appendix F that PPC’s gains transfer when GRPO is replaced with DAPO across all methods.

Generation. For both training rollouts and evaluation we sample with temperature $T = 1.0$ and top-

$p = 0.95$. All experiments are run on $4 \times$ NVIDIA RTX PRO 6000 Blackwell (96GB) GPUs.

4.2 Experimental Results

We organize our evaluation around five questions: whether PPC improves accuracy over plan-based baselines (§4.2.1), whether each reward component contributes (§4.2.2), whether the preplan is genuinely used rather than merely produced (§4.2.3), whether the extra preplan stage inflates inference cost (§4.2.4), and whether PPC’s gain is attributable to problem-specific preplanning rather than structured decomposition (§4.2.5).

4.2.1 Main Results

We evaluate PPC against five baselines: Base, Prompt-Only, GRPO, PTA-GRPO, and Plan-Tuning on five mathematical reasoning benchmarks (AIME25, Minerva-Math, OlympiadBench, MATH-500, GSM8K) across four backbones spanning different families and scales, reporting both *maj@16* and *pass@16* (Table 2).

PPC achieves the best results on 39 of 40 metrics, with consistent gains across all backbones. The improvements are largest on competition-level benchmarks: on Llama3.1-8B, PPC raises MATH-500 *maj@16* by +10.6 over Base and OlympiadBench *pass@16* from 41.3 to 51.4, while on the near-saturated GSM8K, all methods cluster within a narrow band. The Prompt-Only baseline, which prompts the untrained backbone to first analyze *what to solve* before planning and executing, yields marginal and inconsistent improvements over Base: it improves some settings but trails plain GRPO on the large majority of metrics and never approaches PPC, indicating that *prompting* for the preplan cannot substitute for *learning* it. Notably, the prior plan-based methods do not consistently beat plain GRPO on several Minerva-Math and OlympiadBench settings, and occasionally fall below Base. Since PPC differs from these baselines by the upstream preplan, its consistent advantage isolates the effect of modeling *what to solve* before *how to solve*, confirming the paradigm-level gap identified in our paper.

4.2.2 Ablation Study

We ablate the reward components by starting from the outcome reward R_{out} alone and cumulatively adding the spoiler-style penalty R_{sty} , the adherence reward R_{adh} , evaluating each configuration on all five benchmarks across four backbones (Table 3). Note that this ablation is cumulative: the starting

Model	Method	AIME25		Minerva-Math		OlympiadBench		MATH-500		GSM8K	
		maj@16	pass@16	maj@16	pass@16	maj@16	pass@16	maj@16	pass@16	maj@16	pass@16
Qwen3-4B	Base	60.00	73.33	43.01	55.15	66.04	76.81	96.00	98.20	94.84	96.89
	Prompt-Only	60.00	76.67	43.75	54.41	66.48	77.25	96.40	98.20	94.62	97.27
	GRPO	60.00	<u>76.67</u>	44.48	55.88	64.62	<u>77.91</u>	<u>96.60</u>	<u>98.80</u>	94.92	97.04
	PTA-GRPO	60.00	<u>76.67</u>	<u>45.22</u>	<u>58.09</u>	59.89	76.15	95.80	98.60	95.30	<u>97.57</u>
	Plan-Tuning	56.67	<u>76.67</u>	44.85	56.25	66.04	77.25	96.20	98.80	95.00	97.12
	PPC	63.33	80.00	46.32	60.66	67.03	78.35	97.20	99.40	95.15	98.03
Qwen2.5-7B	Base	10.00	30.00	40.07	54.78	40.11	57.47	80.60	89.00	93.71	97.04
	Prompt-Only	16.67	30.00	39.71	58.09	40.77	61.32	80.40	90.00	94.47	97.12
	GRPO	<u>16.67</u>	<u>30.00</u>	41.18	<u>61.76</u>	<u>43.19</u>	59.56	83.80	<u>90.80</u>	94.31	<u>97.57</u>
	PTA-GRPO	16.67	23.33	<u>43.01</u>	58.82	40.77	<u>62.09</u>	81.40	89.40	<u>94.62</u>	97.27
	Plan-Tuning	20.00	26.67	41.54	55.88	41.53	58.13	<u>84.00</u>	88.80	94.09	97.19
	PPC	20.00	36.67	44.85	62.87	44.73	63.85	84.80	92.20	95.22	98.48
Qwen2.5-Math-7B	Base	<u>20.00</u>	30.00	40.81	51.84	43.41	53.96	87.00	90.80	<u>96.97</u>	98.10
	Prompt-Only	16.67	30.00	41.92	52.21	44.62	64.51	87.60	91.80	97.04	98.18
	GRPO	20.00	33.33	42.65	53.31	45.05	64.18	86.80	93.40	96.36	98.40
	PTA-GRPO	16.67	23.33	37.87	<u>56.62</u>	43.52	61.54	<u>87.80</u>	92.80	95.15	98.03
	Plan-Tuning	16.67	26.67	40.07	52.94	43.95	<u>64.73</u>	87.60	<u>93.60</u>	96.82	98.33
	PPC	23.33	36.67	43.38	58.09	46.04	66.26	88.60	95.00	97.12	98.56
Llama3.1-8B	Base	0.00	3.33	35.29	<u>60.29</u>	21.10	41.32	55.20	82.00	89.76	95.98
	Prompt-Only	0.00	3.33	34.56	59.56	23.08	42.42	58.00	80.00	90.37	96.47
	GRPO	0.00	10.00	36.40	59.19	<u>25.05</u>	42.20	59.60	<u>83.40</u>	90.59	96.36
	PTA-GRPO	6.67	10.00	<u>37.13</u>	59.93	24.40	<u>47.47</u>	<u>60.60</u>	82.80	90.75	<u>96.59</u>
	Plan-Tuning	0.00	<u>13.33</u>	36.03	58.09	22.64	47.14	60.20	83.00	<u>90.83</u>	96.13
	PPC	6.67	20.00	37.50	61.40	27.80	51.43	65.80	85.00	92.04	98.03

Table 2: Main results on five mathematical reasoning benchmarks across four backbones. We report maj@16 (self-consistency majority vote) and pass@16 for each benchmark. Within each backbone group and column, **bold** marks the best result and underline the second best. Rows of our method (**PPC**) are highlighted in light blue.

Model	Method	AIME25		Minerva-Math		OlympiadBench		MATH-500		GSM8K	
		maj@16	pass@16	maj@16	pass@16	maj@16	pass@16	maj@16	pass@16	maj@16	pass@16
Qwen3-4B	R_{out} only	46.67	66.67	44.85	57.35	64.40	75.05	93.40	98.20	94.01	97.27
	+ R_{sty}	53.33	70.00	45.22	58.46	65.16	77.58	95.60	98.60	94.54	97.34
	+ R_{adh}	60.00	76.67	45.96	59.56	66.37	78.02	96.60	98.80	94.92	97.80
	PPC (full)	63.33	80.00	46.32	60.66	67.03	78.35	97.20	99.40	95.15	98.03
Qwen2.5-7B	R_{out} only	10.00	23.33	42.28	58.82	41.98	62.20	82.20	91.00	94.39	97.57
	+ R_{sty}	13.33	26.67	42.87	60.29	43.38	62.96	83.40	91.60	94.47	97.95
	+ R_{adh}	16.67	30.00	44.12	61.03	43.40	63.07	84.00	91.80	94.84	98.10
	PPC (full)	20.00	36.67	44.85	62.87	44.73	63.85	84.80	92.20	95.22	98.48
Qwen2.5-Math-7B	R_{out} only	16.67	23.33	41.18	56.25	42.42	63.41	86.40	93.40	96.74	98.10
	+ R_{sty}	20.00	26.67	41.91	56.62	43.08	64.39	87.40	94.20	96.89	98.18
	+ R_{adh}	20.00	33.33	42.27	57.35	44.29	65.05	87.80	94.60	96.97	98.33
	PPC (full)	23.33	36.67	43.38	58.09	46.04	66.26	88.60	95.00	97.12	98.56
Llama3.1-8B	R_{out} only	0.00	10.00	35.66	58.82	23.74	48.35	63.80	82.80	90.98	97.12
	+ R_{sty}	3.33	10.00	36.03	60.29	25.27	49.23	64.60	83.60	91.35	97.19
	+ R_{adh}	6.67	16.67	36.76	61.03	26.37	51.10	65.00	84.80	91.73	97.42
	PPC (full)	6.67	20.00	37.50	61.40	27.80	51.43	65.80	85.00	92.04	98.03

Table 3: Ablation over reward components, starting from R_{out} and cumulatively adding R_{sty} , R_{adh} . R_{fmt} is kept across all settings. Each row adds one term to the row above. We report maj@16 and pass@16; the best result within each backbone group is in **bold**, and our full method (**PPC**) rows are highlighted in light blue.

point “ R_{out} only” uses a binary 0/1 outcome reward, and partial credit (Eq. 7) is introduced only in the final PPC (full) row; thus the + R_{adh} and PPC (full) rows differ solely in whether partial credit is used, cleanly isolating its contribution. The format guard R_{fmt} is kept across all settings.

Every term improves performance monotonically across all four backbones, indicating that the four terms are complementary rather than redundant. Adding R_{sty} alone already lifts AIME25 maj@16 by +6.7 on Qwen3-4B, showing that keeping the

preplan conceptual integrity is beneficial in itself and must be actively maintained during RL, not merely inherited from initialization. The adherence reward contributes its largest increments: a further +6.7 AIME25 maj@16 on Qwen3-4B and +1.9 OlympiadBench pass@16 on Llama3.1-8B: while its effect on the near-saturated GSM8K is marginal, supporting our claim that the preplan pays off only when the plan genuinely follows it. The gap between R_{out} and full PPC widens on competition-level tasks (up to +16.7 AIME25

Benchmark	Setting	Qwen3-4B		Qwen2.5-7B		Qwen2.5-Math-7B		Llama3.1-8B	
		maj@16	pass@16	maj@16	pass@16	maj@16	pass@16	maj@16	pass@16
MATH-500	Shuffled	96.40	98.40	82.60	91.40	87.20	94.00	57.80	80.80
	Mismatched	46.80	79.20	27.80	48.60	36.40	62.80	37.00	67.80
	Generic	94.20	96.80	80.20	90.60	86.40	93.60	59.40	82.00
	PPC	97.20	99.40	84.80	92.20	88.60	95.00	65.80	85.00
AIME25	Shuffled	56.67	76.67	16.67	30.00	16.67	26.67	3.33	6.67
	Mismatched	36.67	56.67	6.67	23.33	3.33	20.00	0.00	0.00
	Generic	56.67	73.33	13.33	26.67	16.67	30.00	3.33	6.67
	PPC	63.33	80.00	20.00	36.67	23.33	36.67	6.67	20.00
Minerva-Math	Shuffled	44.12	59.19	42.27	57.35	42.65	55.14	34.93	56.99
	Mismatched	29.42	45.95	25.37	36.76	26.47	37.13	19.12	37.50
	Generic	41.54	55.88	40.07	55.51	41.17	51.47	36.02	58.46
	PPC	46.32	60.66	44.85	62.87	43.38	58.09	37.50	61.40
GSM8K	Shuffled	92.87	96.97	91.28	96.29	90.14	95.22	90.22	95.90
	Mismatched	78.77	90.67	76.72	85.06	78.99	88.09	71.57	89.00
	Generic	91.73	94.54	90.97	94.99	94.13	95.98	84.68	95.14
	PPC	95.15	98.03	95.22	98.48	97.12	98.56	92.04	98.03

Table 4: Faithfulness analysis on MATH-500, AIME25, Minerva-Math, and GSM8K. We compare PPC against three preplan-perturbation baselines: *Shuffled* (reordered preplan steps), *Mismatched* (a preplan from a different problem), and *Generic* (a content-free generic preplan). We report maj@16 and pass@16; within each column and benchmark block the best result is in **bold** and the second best is underlined. Our method (**PPC**) rows are highlighted in light blue.

Backbone	Preplan	Plan	Execute	PPC Total	GRPO	Plan-Tuning	PTA-GRPO
Llama3.1	0.20	0.17	2.04	2.41	2.67	2.55	5.85
Qwen3	0.17	0.16	1.27	1.60	1.80	1.63	4.18
Qwen2.5	0.15	0.17	1.15	1.47	3.02	1.44	9.97
Qwen2.5-Math	0.12	0.11	0.79	1.02	1.08	0.96	1.16

Table 5: Average generated tokens per problem on MATH-500 (in K tokens). For PPC, *Preplan*, *Plan*, and *Execute* are the three stage-wise sub-costs that sum to *Total*; GRPO, Plan-Tuning, and PTA-GRPO totals are shown for reference. Despite the extra preplan stage, PPC’s total is more compact than full CoT on every backbone and far more compact than PTA-GRPO.

maj@16 on Qwen3-4B) and shrinks on GSM8K, mirroring the main-results trend that explicit problem understanding matters most when the solution path is hard to find.

4.2.3 Faithfulness

To test whether the preplan is genuinely used rather than just produced, we intervene on it at inference time and measure the downstream effects. We replace the model-generated preplan with three perturbations while letting the model freely generate the plan and execution from it: *Shuffled* (the same preplan with its sentences reordered), *Mismatched* (a preplan taken from a different problem), and *Generic* (a content-free placeholder). We report maj@16 and pass@16 on MATH-500, AIME25, Minerva-Math, and GSM8K across four backbones (Table 4).

Injecting a mismatched preplan causes severe collapse on every benchmark and backbone: on Qwen3-4B, AIME25 maj@16 falls from 63.33 to 36.67, Minerva-Math from 46.32 to 29.42, GSM8K from 95.15 to 78.77, and MATH-500

from 97.2 to 46.8; on the weaker Llama3.1-8B, AIME25 collapses entirely to 0.00. This shows that the preplan is genuinely used, not merely produced.

The *Shuffled* and *Generic* conditions jointly confirm that this dependence is on the preplan’s *information*, and this pattern holds consistently across all four benchmarks. A generic, content-free preplan causes only a modest drop (e.g. MATH-500: 94.2 vs. 97.2 on Qwen3-4B; GSM8K: 91.73 vs. 95.15 on Qwen3-4B), serving as a placebo that confirms the collapse under *Mismatched* stems from the injected content, not from the intervention itself. Against this baseline, the contrast with *Shuffled* is telling: when the original information is preserved but its sentences are reordered, accuracy stays close to PPC across benchmarks (e.g. GSM8K 92.87 vs. 95.15 on Qwen3-4B; AIME25 56.67 vs. 63.33 on Qwen3-4B). The model thus relies on *what* the preplan conveys rather than its surface order, confirming that the preplan carries substantive, content-level guidance. Together, these inference-time results show that PPC’s adherence

Qwen3-4B	AIME25		Minerva-Math		MATH-500		GSM8K	
	maj@16	pass@16	maj@16	pass@16	maj@16	pass@16	maj@16	pass@16
Generic Preplan	60.00	73.33	44.85	58.08	96.00	97.80	94.08	97.04
PPC	63.33	80.00	46.32	60.66	97.20	99.40	95.15	98.03
Δ	+3.33	+6.67	+1.47	+2.58	+1.20	+1.60	+1.07	+0.99

Table 6: Training-time control isolating problem-specific preplan content on Qwen3-4B. The generic variant retains structured decomposition, teacher-filtered supervision, and the consistency reward in full, differing from PPC only in whether the preplan analyzes what to solve.

reward (§3.3) drives genuine, content-dependent reliance on the preplan.

4.2.4 Inference Efficiency

To test whether inserting an extra preplan stage inflates inference cost, we measure the average number of generated tokens per problem on MATH-500 for GRPO, Plan-Tuning, PTA-GRPO, and PPC across four backbones (Table 5). All numbers are *total* generated tokens per problem: for PPC this covers all three stages (<preplan> + <plan> + <execute>), and we report this stage-wise breakdown in Table 5 for a fair comparison. The baselines differ in what their supervision teaches the model to emit: GRPO and PTA-GRPO are trained on full DeepSeek-R1 CoT trajectories and therefore inherit long reasoning, whereas Plan-Tuning, like PPC, supervises a structured plan-then-solve format and is correspondingly compact; the length differences thus reflect what each method’s supervision commits the model to producing.

Within PPC’s own budget, the preplan and plan stages together cost only $\sim 0.23\text{--}0.37\text{K}$ tokens across backbones, as both are short by construction: the preplan is a single prose paragraph naming problem type, tools, constraints, and pitfalls, and the plan is a 4–7 step outline that performs no computation. Crucially, the two upstream stages do not simply add length on top of an unchanged execution: the model enters execution having already committed to a solution route, so it advances along a single path instead of exploring and backtracking. Despite producing an additional preplan, PPC’s total is still lower than full CoT on all four backbones (e.g., 1.47k vs. 3.02k on Qwen2.5-7B) and far lower than PTA-GRPO, whose unconstrained reasoning grows to 9.97k tokens on the same backbone; PPC remains within a narrow band of the most compact baseline, Plan-Tuning, while achieving uniformly higher accuracy (§4.2.1).

4.2.5 Isolating Problem-Specific Preplanning

The Generic condition in §4.2.3 perturbs an already-trained model’s preplan at inference time, leaving open whether PPC’s gain instead stems

from structured decomposition, teacher-filtered supervision or the adherence reward alone. To isolate this, we train a controlled variant on Qwen3-4B in which every problem receives the same content-free template preplan (preserving its form and position but removing all what-to-solve information), and the plans are re-generated conditioned on that template; the synthesis pipeline, spoiler filtering, and composite reward are otherwise held fixed. As shown in Table 6, stripping this content costs the most on AIME25 (−6.67 pass@16) and Minerva-Math (−2.58 pass@16), and the least on GSM8K (−0.99 maj@16), with the gap widening as problems get harder. Since this variant retains structured decomposition, teacher-filtered supervision, and the consistency reward in full yet still trails PPC, the gain cannot be attributed to these factors alone; problem-specific preplanning contributes on top of them.

5 Conclusion

We identified a paradigm-level gap in plan-based reasoning: while existing methods model *how to solve* a problem through planning and execution, they leave *what to solve*, the problem type, applicable tools, constraints, and pitfalls, entirely implicit. We proposed **PPC (Preplan-Plan-CoT)**, introducing an explicit preplan stage that gives rise to the *question* $\rightarrow$ *preplan* $\rightarrow$ *plan* $\rightarrow$ *cot* paradigm, realized through a spoiler-score-guarded synthesis pipeline for clean preplan supervision and a composite GRPO reward that enforces the plan genuinely follows from the preplan. Extensive experiments across four backbones and five mathematical reasoning benchmarks demonstrate that PPC achieves the best results on 39 of 40 metrics, improving maj@16 and pass@16 by +2.23 and +3.06 over the strongest baseline at no additional inference-time token cost.

Limitations

Our work has two limitations. First, our framework is evaluated on mathematical reasoning; its effectiveness on other domains such as code gen-

eration or scientific question answering remains unexplored. Second, the spoiler score is a rule-based proxy for the conceptual integrity of preplan, and a more advanced semantic judge could give cleaner preplan supervision.

Acknowledgements

This work is supported by the Guangzhou-HKUST(GZ) Joint Funding Program (No. 2024A03J0630).

References

- Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237.
- Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2026. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. *Advances in neural information processing systems*, 38:110320–110345.
- 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*.
- Zhihao Dou, Qinjian Zhao, Zhongwei Wan, Dinggen Zhang, Weida Wang, Towsif Raiyan, Benteng Chen, Qingtao Pan, Yang Ouyang, Zhiqiang Gao, and 1 others. 2025. Plan then action: High-level planning guidance reinforcement learning for llm reasoning. *arXiv preprint arXiv:2510.01833*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, and 1 others. 2025. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*.
- Runquan Gui, Zhihai Wang, Jie Wang, Chi Ma, Huiling Zhen, Mingxuan Yuan, Jianye Hao, Defu Lian, Enhong Chen, and Feng Wu. 2025. Hypertree planning: Enhancing llm reasoning via hierarchical thinking. *arXiv preprint arXiv:2505.02322*.
- Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. 2025. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850.
- Jujie He, Jiakai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, and 1 others. 2025a. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*.
- Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, and 1 others. 2025b. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *arXiv preprint arXiv:2504.11456*.
- Xingshuai Huang, Derek Li, Bahareh Nikpour, and Parsa Omid. 2026. Hierarchical chain-of-thought prompting: Enhancing llm reasoning performance and efficiency. *arXiv preprint arXiv:2604.00130*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*.
- 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*.

Mihir Parmar, Palash Goyal, Xin Liu, Yiwen Song, Mingyang Ling, Chitta Baral, Hamid Palangi, and Tomas Pfister. 2025. Plan-tuning: Post-training language models to learn step-by-step planning for complex problem solving. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21430–21444.

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*.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 2609–2634.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhao Chen. 2024. Mammoth: Building math generalist models through hybrid instruction tuning. In *International Conference on Learning Representations*, volume 2024, pages 40320–40341.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans,

Claire Cui, Olivier Bousquet, Quoc Le, and 1 others. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Appendix Overview

Appendix	Title
Appendix A	Error Attribution Analysis
Appendix B	Spoiler/Leakage Failure: An Example
Appendix C	Three-Stage Data Generation Prompts
↪ Appendix C.1	Stage 1: Preplan Prompt
↪ Appendix C.2	Stage 2: Plan Prompt
↪ Appendix C.3	Stage 3: Execution Prompt
Appendix D	The Spoiler Score $s(\cdot)$
Appendix E	Reward Details
↪ Appendix E.1	Partial-Credit Utility and Reliability
Appendix F	Robustness to the RL Backbone

A Error Attribution Analysis

To substantiate our claim that methods without an explicit preplan stage fail because misunderstand *what* to solve rather than *how*, we run an LLM-as-judge analysis over each method’s incorrectly answered problems. Rather than inferring the claim indirectly, we let the judge decide it: for each wrong solution, it judges whether the failure stems from not understanding *what to solve*, defined via four facets a short computation-free analysis could catch: *problem type*, *tools/concepts*, *constraints*, and *pitfalls*. The judge prompt is shown in Figure 3.

To verify that findings in Figure 1 is *judge-agnostic* rather than an artifact of a single judge, we repeat the identical attribution pipeline with a Qwen judge. As shown in Figure 4, the result is highly consistent: across all four backbones, PPC again commits substantially fewer *what-to-solve* errors than the baselines, with the same ranking and comparable magnitudes. This confirms that our observation: the gains of pre-planning stem from better understanding *what* to solve is robust to the choice of judge.

B Spoiler/Leakage Failure: A Full Example

This appendix expands on the spoiler-score filter introduced in §3.2. Figure 5 contrasts a *discarded* preplan that crosses the conceptual integrity boundary with a *retained* clean preplan on a problem of comparable difficulty. Both trajectories reach a correct final answer; only the first is discarded, showing that the filter targets preplan purity, not answer correctness.

Error Attribution Prompt

You are an expert mathematician diagnosing WHY a model's solution to a competition math problem is wrong. The model produced
 $\hookrightarrow$ an INCORRECT final answer.

Your single most important job is to decide ONE thing:

>> Was this failure caused by the model not understanding WHAT TO SOLVE?

Definition you MUST use. "Understanding WHAT TO SOLVE" is the non-computational understanding of the problem that should
 $\hookrightarrow$ happen BEFORE any calculation. It has exactly four facets:

- (1) PROBLEM TYPE - recognizing what kind of problem this is and what overall approach it calls for.
- (2) TOOLS/CONCEPTS - knowing which theorem, formula, concept, or technique is the right one to bring to bear.
- (3) CONSTRAINTS - noticing the boundary conditions, domain restrictions, edge cases, or the multiple cases that must be
 $\hookrightarrow$ considered.
- (4) PITFALLS - anticipating well-known traps (misreading a condition, double counting, sign/orientation issues baked
 $\hookrightarrow$ into the setup, off-by-one in the framing).

A "WHAT-TO-SOLVE failure" means the root cause is a breakdown in one of these four facets: the model went wrong because it
 $\hookrightarrow$ misjudged the nature of the problem, brought the wrong tool, ignored a constraint, or walked into a foreseeable trap -- i.e.
 $\hookrightarrow$ a short upfront analysis of the problem (without doing any calculation) would plausibly have caught it.

This is the OPPOSITE of a "HOW-TO-SOLVE failure", where the model correctly understood what the problem needs and chose a
 $\hookrightarrow$ sound approach, but failed while CARRYING IT OUT.

[... full label definitions, decision rules, and JSON output format omitted]

Figure 3: The error attribution prompt (abridged).

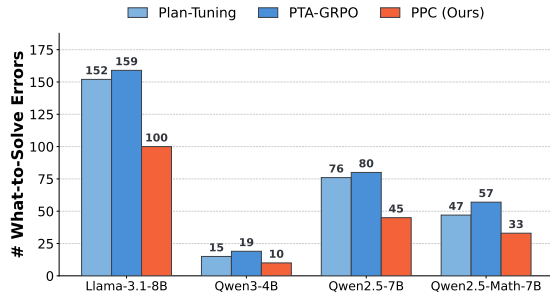


Figure 4: Number of *what-to-solve* errors on MATH-500 across four backbones, attributed by a Qwen judge (cf. Figure 1, judged by DeepSeek-V4). The ranking and magnitudes closely match, indicating the attribution is robust to the choice of judge.

Why the first is discarded. The discarded preplan does not survey *what to solve*; it has already settled *how to solve*. It commits to concrete computed outcomes that both singularities are enclosed and that the origin is an *essential* singularity (a spoiler), and rehearses the exact downstream procedure of expanding both Laurent series, multiplying, and extracting the $1/z$ coefficient (a leakage). These trip the derivation-phrasing and answer-revealing signals of $s(\cdot)$, pushing its score above τ_s despite the correct answer $2\pi i$.

Why the second is retained. The retained preplan stays on the *what-to-solve* side throughout: it names the relevant tool (CRT), notes the structural fact $1001 = 7 \cdot 11 \cdot 13$, and flags genuine pitfalls—without solving any congruence or committing to a step sequence. The actual reconstruction

is left to the planning and execution stages.

C Three-Stage SFT Data Generation Prompts

This appendix provides the full prompts used to construct the three-stage (preplan, plan, execute) trajectories of §3.2. The stages are generated strictly left-to-right, and each stage's prompt is shown below (with the runtime fields {question}, {preplan}, {plan} substituted at generation time).

Stage 1: Preplan. The preplan is produced first, before any plan exist. The prompt (Figure 7) casts the planner LLM as a teacher *briefing* a student rather than solving the problem, and enforces the conceptual integrity through explicit rules: no derivations, no extended formulas, no computed values, and prose rather than steps. These rules operationalize the spoiler constraint that the filter of §3.2 later enforces.

Stage 2: Plan. Conditioned on the question *and* the preplan, the same instruction-tuned LLM expands the high-level direction into a concrete numbered plan (Figure 8). The prompt requires the plan to faithfully follow the strategy hinted at in the preplan, which is what makes the plan a genuine elaboration of the preplan rather than an independent solution.

Stage 3: Execute. The executor sees only the question and the plan, never the preplan, so that execution is driven by the explicit plan alone. We realize execution with a strong reasoning model

Discarded preplan (spoiler + leakage)

Spoiler score $s(y_{pp}) = 4$; final answer correct; discarded.

This is a complex contour integral problem ... using the residue theorem. ... Since the contour is the circle $|z - 1| = \frac{3}{2}$, it **encloses both** $z = 0$ **and** $z = 1$, so we must consider the nature of each singularity and how they contribute to the residue sum. The problem specifies using Laurent series expansions for $e^{1/z}$ and $\frac{1}{z-1}$, which implies that **understanding how to multiply and manipulate these series to extract the coefficient of $\frac{1}{z}$ will be central** to determining the residue at $z = 0$. We should also be careful about distinguishing between removable singularities, poles, and essential singularities ... A key pitfall is misidentifying the order of a pole or mishandling the series expansions, especially near the **essential singularity at the origin**.

Retained preplan (clean)

Spoiler score $s(y_{pp}) = 2 \leq \tau_s$; retained.

This is a modular arithmetic problem involving large exponents and simultaneous congruences. ... The key concepts at play include the Chinese Remainder Theorem and properties of modular exponentiation. Since 1001 factors into the primes 7, 11, and 13, the given congruences suggest that the solution involves combining these individual modular results into a unified modulus. A natural approach is to treat each congruence as a piece of a larger puzzle and reconstruct the final result modulo 1001. We should also be careful not to assume uniqueness without verifying compatibility of the congruences. One potential pitfall is misapplying the theorem or mixing up moduli during calculations ...

Figure 5: Two preplans of comparable difficulty, both reaching correct answers. **Top:** a discarded preplan that *computes and pre-enacts*—asserting computed classifications and rehearsing the solution steps. **Bottom:** a retained preplan that *analyzes and defers*—naming tools and pitfalls without solving. The spoiler score $s(\cdot)$ separates the two.

Computation of the spoiler score $s(\cdot)$

```
Input : preplan text y
Output: spoiler score s in {0,...,6}

DERIVATION_PHRASES = {"simplifies to", "reduces to",
    "leads to", "results in", "yields", "gives us",
    "implies that", "becomes", "transforms to",
    "evaluates to", "rewrites as"}
ANSWER_PHRASES = {"the answer", "the result is",
    "we get", "we obtain"}

s1 = any phrase in DERIVATION_PHRASES occurs in y
s2 = ( count("=", y) + count("\equiv", y) ) >= 3
s3 = number of inline-math spans  $\$....\$$  with
    length >= 30 is >= 2
s4 = y contains a standalone integer with >= 3 digits
s5 = any phrase in ANSWER_PHRASES occurs in y

s6 = number of inline-math spans  $\$....\$$  is >= 4

s = s1 + s2 + s3 + s4 + s5 + s6
return s      # retain trajectory iff s <= tau_s
```

Figure 6: Rule-based computation of the spoiler score. Each of the six signals contributes one point; a preplan is retained only when $s \leq \tau_s$ (we use $\tau_s = 2$).

(DeepSeek-R1) that produces a raw worked solution, which is then reorganized by the instruction-tuned model into a clean numbered execution whose step titles align one-to-one with the plan (Figure 9). The cleanup step preserves all numerical results from the raw solution rather than re-deriving them, keeping the mathematics intact while imposing a consistent structure.

Verification. A trajectory is kept only if its final answer matches the gold answer (verified by normalized string match with an LLM equivalence check as fallback) and its execution ends in the required `\boxed{}` format.

D The Spoiler Score $s(\cdot)$

This appendix specifies the rule-based spoiler score $s(\cdot)$ used by the filter in Eq. 4. The score is an integer in $\{0, \dots, 6\}$ obtained by summing six binary signals, each of which fires when the preplan shows a symptom of *computing* or *revealing the answer* rather than *analyzing* the problem. A preplan is retained only when $s(y_{pp}) \leq \tau_s$; we use $\tau_s = 2$. The score is applied uniformly at construction time (Eq. 4) and as a reward component during RL without inconsistency.

The six signals. Let y_{pp} be the candidate preplan text. Each signal below contributes 1 to $s(y_{pp})$ when its condition holds:

- Derivation phrasing.** The text contains any phrase signalling an inference or transformation has been carried out: e.g. “simplifies to”, “reduces to”, “leads to”, “yields”, “implies that”, “becomes”. Such phrases indicate the preplan is performing, not previewing, the derivation.
- Equation density.** The combined count of equality symbols (= and `\equiv`) is at least

Stage 1 — Preplan Prompt

You are a math teacher briefing a student before they attempt a problem. You yourself MUST NOT solve the problem. Your only $\hookrightarrow$ job is to set the stage so that the student knows HOW TO APPROACH the problem mentally before starting.

```
<question>
{question}
</question>
```

Write a brief pre-solution analysis as a single coherent paragraph (around 4-8 sentences). Cover the following aspects

$\hookrightarrow$ naturally -- without using labels or section headers:

- What TYPE of problem is this? (e.g., "this is a contour integral", "this is a divisibility problem")
- What general TOOLS or CONCEPTS are likely useful? (Name them, do not apply them.)
- What is the high-level STRATEGIC DIRECTION? (Describe in plain words, not as steps.)
- What CONSTRAINTS or boundary conditions matter?
- What PITFALLS should be anticipated?

ABSOLUTE RULES (any violation invalidates your output):

1. NO derivations. Do NOT use phrases like "this simplifies to", "this leads to", "this implies", "this becomes", "we get", " $\hookrightarrow$ the result is".
2. NO formulas longer than 15 characters. You may name a quantity (e.g., "the integral", "the polynomial $f(x)$ ") but DO NOT $\hookrightarrow$ write extended expressions in $\$....\$$.
3. NO equations beyond definitional ones (e.g., "let n be the count of...").
4. NO specific computed values. Even if you know the answer, do NOT mention it.
5. NO step-by-step procedures. The plan comes later -- your job is meta-thinking.
6. Write as flowing prose, NOT as a list. Use natural transitions like "this suggests", "a natural approach is", "we should $\hookrightarrow$ also be careful about".
7. Be reasonably concise -- focus on insights, not exhaustive coverage.

Output the paragraph directly. No headers, no labels, no quotation marks around the output.

Figure 7: Stage 1 (preplan) prompt, which constrains the output to high-level analysis rather than step-by-step solving

Stage 2 — Plan Prompt

You are given a math problem and a pre-analysis written by a teacher. Your job is to translate the high-level strategic $\hookrightarrow$ direction in the pre-analysis into a concrete numbered solution plan.

```
<question>
{question}
</question>
```

```
<pre_analysis>
{preplan}
</pre_analysis>
```

Create a numbered plan (4-7 steps). Each step should:

- Have a brief title and a one-sentence description
- Describe WHAT to do and WHY (referencing the strategy from the pre-analysis)
- NOT perform actual calculations (that comes in execution)

The plan MUST faithfully follow the strategy hinted at in the pre-analysis. Do NOT invent a different approach.

Output EXACTLY in this format (no other text):

1. [Step Title]: [one sentence describing what to do and why]
2. [Step Title]: [one sentence]
3. ...

Rules:

- Each step $\leq$ 2 sentences.
- NO specific numerical computations.
- NO LaTeX formulas longer than 20 characters.
- The total plan should be 400-1000 characters.

Figure 8: Stage 2 (plan) prompt. The plan is conditioned on the preplan and must faithfully follow its strategy.

three. A preplan that states constraints needs few equalities; a high count signals worked relations.

3. **Long inline expressions.** At least two inline-math spans exceed 30 characters. Short named quantities are permitted, but extended expressions indicate concrete manipulation.

Stage 3 — Execution Cleanup Prompt

You are given a mathematician's raw solution work for a math problem, and the solution plan that was supposed to guide it.
⇒ Your task is to organize the raw solution into a clean, structured execution.

```
<question>
{question}
</question>

<solution_plan>
{plan}
</solution_plan>

<raw_solution>
{raw_solution}
</raw_solution>
```

Organize the raw solution into a structured numbered execution.

Rules:

1. Each step title MUST exactly match the corresponding plan step title.
2. Include the full mathematical reasoning and calculations from the raw solution.
3. If the raw solution contains errors, self-corrections, or multiple attempts, use the FINAL corrected version.
4. Preserve ALL numerical calculations from the raw solution. Do NOT re-derive.

Your output MUST end with exactly: Final Answer: `\boxed{answer}`

Output EXACTLY in this format (no other text before or after):

```
1. [Title matching plan step 1]: [calculation and reasoning]
2. [Title matching plan step 2]: [calculation and reasoning]
...
Final Answer: \boxed{answer}
```

Figure 9: Stage 3 cleanup prompt. The raw reasoning trace is reorganized so that step titles align with the plan, preserving (not re-deriving) all computed values.

Plan-adherence judge prompt (R_{adh})

You are evaluating whether a solution plan truly follows from a pre-analysis (preplan).

```
Question: {question}
Pre-analysis (preplan): {preplan}
Solution plan: {plan}
```

Rate how well the plan FOLLOWS the preplan's strategic direction (1-5):

- 5: Plan tightly implements the strategy hinted in preplan; the tools/concepts mentioned in preplan appear in the plan steps;
⇒ the plan would be DIFFERENT if the preplan suggested a different approach.
- 4: Mostly follows, with minor unrelated additions or one missing element.
- 3: Partially aligns; some strategic elements reflected, but the plan also wanders into directions the preplan did not hint
⇒ at.
- 2: Only loosely connects; the plan would look largely the same even under a different preplan.
- 1: Completely ignores or contradicts the preplan.

IMPORTANT: Score by STRATEGY ALIGNMENT, not by quality. A correct plan that ignores the preplan should still score LOW; an OK

⇒ plan that faithfully follows it should score HIGH.
Output ONLY a single integer 1-5. No explanation.

Figure 10: The frozen plan-adherence judge prompt for R_{adh} .

4. **Multi-digit constants.** The text contains a standalone integer of three or more digits, which typically appears only once specific numbers are being computed or stated.
 5. **Answer-revealing wording.** The text contains phrases such as “the answer”, “the result is”, “we get”, or “we obtain”, which tend to precede a disclosed outcome.
 6. **Inline-math span count.** The total number of inline-math spans is at least four, indicating the preplan leans on formal expressions rather than prose analysis.
- Aggregation and rationale.** The final score sums the six indicators,

Method	AIME25		Minerva-Math		MATH-500		GSM8K	
	maj@16	pass@16	maj@16	pass@16	maj@16	pass@16	maj@16	pass@16
DAPO	60.00	80.00	45.22	55.88	96.80	98.60	95.07	97.27
PTA-DAPO	63.33	80.00	45.95	58.82	96.20	98.80	95.68	97.35
Plan-Tuning	63.33	76.67	45.58	57.72	96.20	99.00	95.22	97.12
PPC	63.33	83.33	47.43	61.03	97.40	99.20	95.83	98.18

Table 7: Robustness of PPC to the RL backbone: full comparison on Qwen3-4B with DAPO substituted for GRPO in every method, keeping the backbone matched across methods. PPC remains best or tied-best on every metric, confirming the gains stem from the preplan paradigm rather than the choice of RL algorithm.

$$s(y_{pp}) = \sum_{i=1}^6 \mathbb{I}[\text{signal}_i \text{ fires on } y_{pp}] \in \{0, \dots, 6\}, \quad (8)$$

This additive design tolerates the incidental use of light notation that even a purely analytical preplan may contain, and targets the *accumulation* of computational behaviour that characterizes a true spoiler. The procedure is summarized in Figure 6.

E Reward Details

In this section, we show the details of each reward component and exam partial credit’s utility and reliability. All judges are *frozen* (a fixed instruction-tuned LLM, not updated during RL), so the reward signal is stationary across training.

Outcome (R_{out}). To provide a gradient beyond a binary signal, so that GRPO can still learn once most rollouts are wrong, an incorrect answer is scored by a *proximity* judge on a 1–5 scale (mapped to $[0 - 0.5]$) that rates how close the *solution path* was to a correct one, deliberately ignoring numerical closeness of the final value.

Format (R_{fmt}). A rule-based check awards 0.3 iff the completion has a clean structure: the three tags `<preplan>`, `<plan>`, `<execute>` each appear exactly once, in order, with a `\boxed{\}` answer inside `<execute>`, no trailing text after `</execute>`, and no tags outside the allowed set.

Style (R_{sty}). To keep the preplan from regressing into a spoiler, we reuse the spoiler score $s(\cdot)$ of Appendix D as a one-sided penalty $R_{\text{sty}} = \max(0, s(y_{pp}) - \tau_s)$ with $\tau_s = 2$, weighted by $\lambda_s = 0.1$ in Eq. (6). Thus the penalty term $-\lambda_s R_{\text{sty}}$ is 0 while $s \leq 2$ and decreases (down to -0.4 when s reaches its maximum of 6) as the preplan accumulates computational signals.

Adherence (R_{adh}). A frozen judge rates, on a 1–5 scale, whether the plan implements the *strategy* of the preplan (not its quality), scaled to a maximum of 0.1. Crucially, a plan that solves the problem well but ignores the preplan still scores low, which

is what ties the plan to the preplan rather than letting it become an independent solution. The full prompt is shown in Figure 10.

Partial-credit utility and reliability. We instrumented GRPO to log every rollout’s verdict and outcome reward (Qwen3-4B). In 70.0% of groups, incorrect rollouts received distinct partial-credit values, and in groups where all 8 rollouts are wrong, partial credit restores a usable gradient in 92.2% of cases, converting undifferentiated incorrect rollouts into a graded, learnable signal. We further verified the reward scale’s consistency on the actual scored rollouts collected during training: re-scoring 150 incorrect trajectories three times with the same judge (Qwen) yields identical bands in 98.0% of cases, with an induced reward fluctuation averaging $0.004 - 0.8\%$ of the $[0, 0.5]$ partial-credit range; re-scoring with an independent judge (DeepSeek-V4) shows that for 86.0% of trajectories the two judges’ 1–5 ratings are identical or adjacent, confirming stable reward assignment both within and across judges.

F Robustness to the RL Backbone

The standard implementations of the plan-based baselines we compare against, Plan-Tuning and PTA-GRPO, are both built on GRPO, so we adopt it keeps the RL backbone identical across all methods and isolates the effect of the preplan stage.

To verify that our gains do not depend on this specific choice of RL algorithm, we re-ran the full comparison on Qwen3-4B with DAPO substituted for GRPO in every method, so the backbone remains matched across methods. As shown in Table 7, PPC remains the best or tied-best on every metric under DAPO. This confirms that the gains come from the preplan paradigm rather than from any property of GRPO specifically, and that PPC transfers across policy-optimization algorithms.