

ATTNPO: Attention-Guided Process Supervision for Efficient Reasoning

Shuaiyi Nie^{1,2}, Siyu Ding^{3†}, Wenyuan Zhang^{1,2}, Linhao Yu⁴, Tianmeng Yang³, Yao Chen^{1,2},
Weichong Yin³, Yu Sun³, Hua Wu³, Tingwen Liu^{1,2†}

¹Institute of Information Engineering, Chinese Academy of Sciences

²School of Cyber Security, University of Chinese Academy of Sciences

³Baidu Inc. ⁴Tianjin University

{nieshuaiyi, liutingwen}@iie.ac.cn

Abstract

Large reasoning models trained with reinforcement learning and verifiable rewards (RLVR) achieve strong performance on complex reasoning tasks, yet often overthink, generating redundant reasoning without performance gains. Existing trajectory-level length penalties often fail to effectively shorten reasoning length and degrade accuracy, as they uniformly treat all reasoning steps and lack fine-grained signals to distinguish redundancy from necessity. Meanwhile, process-supervised methods are typically resource-intensive and suffer from inaccurate credit assignment. To address these issues, we propose ATTNPO, a low-overhead process-supervised RL framework that leverages the model’s intrinsic attention signals for step-level credit assignment. We first identify a set of special attention heads that naturally focus on essential steps while suppressing redundant ones. By leveraging the attention scores of these heads, we then employ two sub-strategies to mitigate overthinking by discouraging redundant steps while preserving accuracy by reducing penalties on essential steps. Experimental results show that ATTNPO substantially reduces reasoning length while significantly improving performance across 9 benchmarks¹.

1 Introduction

Recent advances in Large Reasoning Models (LRMs) (Guo et al., 2025; OpenAI, 2024) have demonstrated remarkable performance on complex reasoning tasks. This success is largely driven by outcome-supervised RL methods, represented by GRPO (Shao et al., 2024), which omits the critic model in PPO (Schulman et al., 2017) by utilizing verifiable outcome rewards to estimate group advantages. A hallmark of LRMs is their ability to produce deliberative long CoT that employ advanced strategies such as self-reflection and multi-

¹Code link: <https://github.com/NieSYsc20/AttnPO>.

[†]Corresponding author. [‡]Project lead.

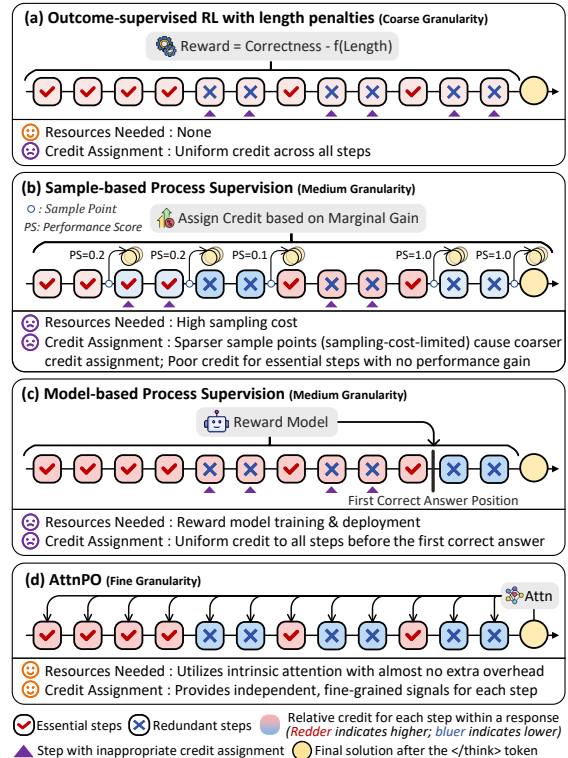


Figure 1: ATTNPO vs. other reinforcement learning methods for efficient reasoning.

path exploration (Li et al., 2025b). However, long CoT fosters overthinking (Zhang et al., 2025b): LRMs indiscriminately apply verbose reasoning, wasting computation on even trivial operations.

Integrating length penalties in outcome-supervised RL (Figure 1(a)) is widely adopted to mitigate overthinking. The core idea is to assign higher rewards to responses that are both correct and concise (Arora and Zanette, 2025; Team et al., 2025; Yi et al., 2025). However, this coarse-grained feedback assigns uniform credit to all reasoning steps, failing to provide discriminative learning signals to distinguish between necessary and redundant steps.

Therefore, recent studies focus on process supervision to down-weight the credit assigned to redun-

dant steps. As shown in Figure 1, sample-based methods (Figure 1(b)) estimate the marginal performance gain of intermediate steps via Monte Carlo sampling of immediate final solutions (Dai et al., 2025; Yue et al., 2025), while model-based methods (Figure 1(c)) train a reward model to pinpoint the first correct answer position, treating only the subsequent steps as redundant (Jiang et al., 2025; Tan et al., 2025b). However, these methods face two key limitations: (1) High Overhead, which requires additional sampling or a reward model; (2) Inaccurate Credit Assignment, which fails to provide independent step-wise signals and may lead to erroneous credit allocation. A key question thus emerges: **Can we, with almost no additional resource costs, rely solely on the model’s intrinsic signals to achieve fine-grained, step-wise supervision of which steps are essential or redundant?**

In this study, we delve into the model’s intrinsic attention mechanism and identify a striking phenomenon: during the final answer generation (following the `</think>` token), a subset of heads naturally focus on essential steps by assigning them higher attention weights while suppressing redundant ones. We term these **Key-Focus Heads** (KFHs). Building on this finding, our experiments further reveal: (1) Positional Distribution, KFHs are mainly located in middle-to-late layers; (2) Training Stability, where the behavior of KFHs varies little under RL training with length penalties; and (3) Sparse Sufficiency, as a small set of KFHs suffice for distinction, with performance saturating rapidly as more heads are added.

Based on these insights, we propose **ATTNPO** (Attention-guided Policy Optimization), a novel RL framework that enables step-level supervision by leveraging KFHs’ attention scores, achieving superior performance with negligible training overhead. Specifically, ATTNPO scales the outcome-level advantage in a stepwise manner via two strategies on correct responses: (1) for correct responses with positive outcome advantages, ATTNPO attenuates the positive advantage assigned to redundant steps, preventing over-encouragement them and mitigating overthinking; and (2) for correct responses with negative outcome advantages, ATTNPO attenuates the negative advantage assigned to essential steps, avoiding excessive penalization of valid reasoning and alleviating performance degradation.

Our contributions are threefold: (1) We are the first to reveal the existence of KFHs in LRMs, which naturally focus on essential reasoning steps

while suppressing redundant ones; (2) Based on this insight, we propose ATTNPO, a low-overhead process-supervised RL framework that leverages intrinsic attention signals for step-level credit assignment; (3) Empirical results demonstrate that ATTNPO substantially shortens reasoning length while improving accuracy. For example, on DeepSeek-R1-Distill-Qwen-1.5B, ATTNPO achieves an average +7.3-point absolute accuracy gain with a 60% reduction in reasoning length across six math benchmarks.

2 Preliminary

Two-stage Generation of LRMs Given an input question q , an LRM \mathcal{M} generates a response $o = (\mathcal{T}, \mathcal{F})$ in two sequential stages, explicitly separated by a special `<think>` token. First, the model produces a deliberate thinking process \mathcal{T} , which may contain redundant reasoning structures, such as unnecessary reflection or verification (Jiang et al., 2026). Then, it outputs a concise final solution \mathcal{F} by summarizing \mathcal{T} and retaining only the essential steps required to solve the problem.

Step Segment The thinking process \mathcal{T} can be decomposed into meaningful intermediate steps, whose boundaries are typically marked by specific special phrases that capture key points in reasoning, such as pauses, reflections or shifts in direction (Hammoud et al., 2025; Choi et al., 2025; Yue et al., 2025). Thus we segment the thinking process with special phrases, which fall into three categories: Confused phrases (e.g., “Wait”) indicate the model is entering a reflection, verification, or branching stage; Progression phrases (e.g., “First”) signal the continuation of the current reasoning; Summary phrases (e.g., “Therefore”) summarize the preceding reasoning. To avoid overly short segments, we adaptively merge segments shorter than a threshold. During segmentation and merging, we separate segments that begin with confusion phrases from the preceding context, since they often trigger redundant verification or reflection. Isolating them makes such redundancy easier to detect and penalize. See Appendix A.1 for details.

3 Key-Focus Heads

Following the two-stage generation paradigm described in 2, LRMs transform a deliberative thinking process—often containing redundant reasoning—into a concise final solution, which requires non-trivial selection and aggregation of essential

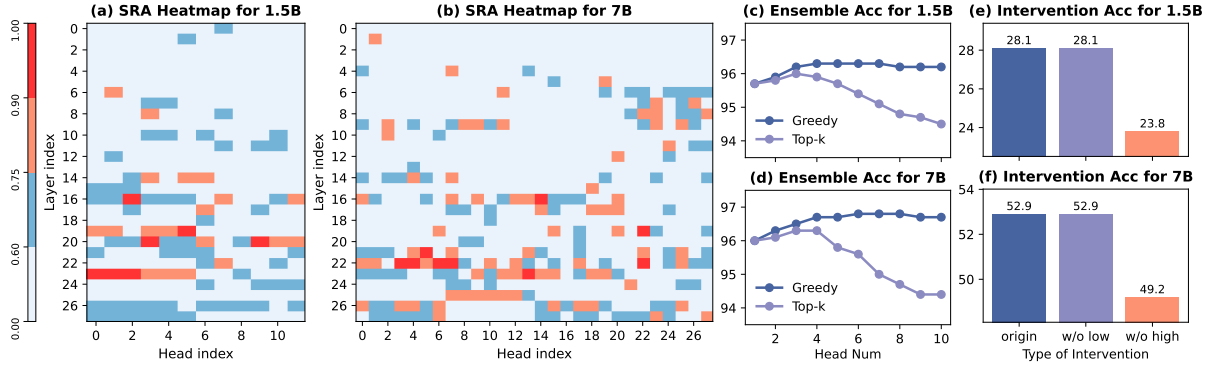


Figure 2: Probing results of Key-Focus Heads.

information. In Transformer architecture, attention serves as the primary mechanism for information selection (Vaswani et al., 2017), and prior work has shown that different attention heads specialize in distinct functions (Zheng et al., 2024; Li et al., 2025a; Chen et al., 2026). Based on this observation, we hypothesize that **during final-answer generation, there exist a specific subset of attention heads that selectively focus on essential reasoning steps while ignoring redundant ones, thereby summarizing and integrating the thinking process.** To test this hypothesis, we examine the attention distributions over essential and redundant steps across different attention heads.

3.1 Probing Setup

Probe Data Construction To analyze how different attention heads attend to essential versus redundant steps, we require a probing dataset with reliable step-level annotations. We therefore sample 300 questions from Deepscaler (Luo et al., 2025b) that are not difficult for DeepSeek-R1-Distill-Qwen-1.5B (average accuracy ≥ 0.875 over eight samples). This restriction is important because, on harder problems, LRMs often exhibit exploratory reasoning. Such reasoning may appear redundant, but still contributes to correct solutions, making reliable step annotation difficult and thereby affecting the head identification. For each question, we sample eight responses, select the median-length correct one, and segment its thinking process following Section A.1. We then annotate each reasoning step using three state-of-the-art LLMs², and retain only the steps on which all three models agree on the label, thereby reducing annotation noise. Thus, for each thinking process \mathcal{T} , we obtain a set of essential steps E_s and a set of redundant steps R_s . Details are provided in A.2.1.

²We use GPT-5, Claude-4-sonnet, and Gemini-2.5-pro.

Attention-Based Step Scores For each head at layer l and head h , we compute an attention-based step score $\mathcal{S}_{s_k}^{l,h}$ for each reasoning step s_k , where k indexes the step index, based on the attention weights used during final-answer generation. Specifically, let $a_{m \rightarrow n}^{l,h}$ denote the attention weight from token m in the final solution \mathcal{F} to token n in step s_k . We define the step score $\mathcal{S}_{s_k}^{l,h}$ as:

$$\mathcal{S}_{s_k}^{l,h} = \frac{1}{|s_k|} \sum_{m \in \mathcal{F}} \sum_{n \in s_k} a_{m \rightarrow n}^{l,h}. \quad (1)$$

This score measures the average attention received by each token in step s_k from the final solution.

Evaluation Metrics: Step Ranking Accuracy

We evaluate how well each attention head distinguishes between essential and redundant steps. For each thinking process \mathcal{T} , we form all essential–redundant step pairs $\mathcal{P}_{\mathcal{T}} = \{(s^+, s^-) \mid s^+ \in E_s, s^- \in R_s\}$. For head h at layer l , we define Step Ranking Accuracy (SRA) as:

$$\text{SRA}_{l,h} = \frac{|\{(s^+, s^-) \in \mathcal{P}_{\mathcal{T}} \mid \mathcal{S}_{s^+}^{l,h} > \mathcal{S}_{s^-}^{l,h}\}|}{|\mathcal{P}_{\mathcal{T}}|}. \quad (2)$$

SRA measures the fraction of step pairs in which the essential step receives a higher score than the redundant one. We report the final SRA by averaging across all examples in the probing dataset.

3.2 Probing Analysis

We conduct experiments on DeepSeek-R1-Distill-Qwen-1.5B / 7B and make the following observations; more detailed results are in Appendix A.2.2:

- *During final solution generation, some attention heads selectively focus on essential reasoning steps while ignoring redundant ones, which are mainly located in the middle-to-late layers.* We refer to these specialized heads as **Key-Focus Heads**. As shown in Figure 2(a,b), across both

evaluated LRMs, a small number of heads achieve SRA above 0.9, with the best-performing head reaching 95% in the 1.5B model and 96% in the 7B model. Most heads, by contrast, perform at or below the random baseline (SRA ≈ 0.5).

- *Ensembling step scores \mathcal{S}_{s_k} from multiple attention heads provides only limited gains on the probing dataset.* In Figure 2(c,d), under both *top-k selection by SRA* and *greedy SRA-maximizing selection*, performance improves slightly when a small number of heads are combined, but quickly saturates or degrades as more heads are added.
- *Under RL training with length penalties, the behavior of KFHS varies little, suggesting that their functional roles are robust.* We train the model using the reward function described in Eq. 5 and re-run the same probing process on checkpoints at steps 500 and 1000. The resulting SRA distributions show high consistency across checkpoints, with pearson correlation coefficients exceeding 0.85 in all cases.
- *Even when using non-difficult question in probing, KFHS exhibit a degree of generalization to challenging problems.* We sample responses on the challenging AIME24 dataset and obtain step scores from the head with the highest SRA. From each thinking process \mathcal{T} with a correct final answer, we remove either the bottom 30% (low-scoring) or the top 30% (high-scoring) steps, and then append the prompt “*I think I have finished thinking. Now give the final solution step by step.*” to generate the final solution. As shown in Figure 2(e,f), removing high-scoring steps degrades pass@32, while removing low-scoring steps has little impact.

4 ATTNPO

In this section, we propose ATTNPO (Attention-guided Policy Optimization), which leverages KFHS to distinguish redundant from essential steps, thereby enabling fine-grained stepwise supervision without additional resources. We first introduce the overall stepwise advantage rescaling framework in Section 4.1, followed by two complementary strategies under this framework in Sections 4.2 and 4.3. The overall framework is illustrated in Figure 3.

4.1 Stepwise Advantage Rescaling

Current outcome-supervised RL methods for efficient reasoning typically adopt a response-level reward formulation that integrates correctness with

length penalties. The resulting rewards are then leveraged to estimate a value baseline, thereby yielding relative advantages for gradient updates, as exemplified by GRPO (Shao et al., 2024) and RLOO (Ahmadian et al., 2024). Typically, these algorithms share the following objective function:

$$\mathcal{J}(\theta) = \mathbb{E}_{q \sim P, \{o_i\}_{i=1}^G \sim \pi_\theta} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left(\min(\rho_{i,t} A^i, \text{clip}(\rho_{i,t}, 1 - \lambda, 1 + \lambda) A^i) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right) \right]. \quad (3)$$

Here, A^i denotes the outcome-level advantage, and G the number of rollouts per input q . The importance sampling ratio is $\rho_{i,t} = \pi_\theta(o_{i,t} | q, o_{i,<t}) / \pi_{\theta_{\text{old}}}(o_{i,t} | q, o_{i,<t})$. Although these methods achieve improved performance while eliminating the high-overhead critic model used in PPO (Schulman et al., 2017), they suffer from coarse-grained credit assignment: the advantage A^i is uniformly applied to all steps within a rollout, regardless of their varying degrees of redundancy.

Thus, we propose **Stepwise Outcome-level Advantage Rescaling** to rescale the outcome-level advantage A^i at each step s_k of a correct response:

$$\hat{A}_{s_k}^i = \gamma_{s_k}^i \cdot A^i, \text{ where } \gamma_{s_k}^i \geq 0. \quad (4)$$

Here, $\hat{A}_{s_k}^i$ denotes the step-level advantage used for policy updates at step s_k , and $\gamma_{s_k}^i$ is a non-negative scaling factor computed from the attention scores of KFHS, reflecting step-level redundancy, which will be detailed in the following section. This formulation explicitly captures the relative redundancy of individual steps within a single reasoning process. By scaling the advantage according to stepwise redundancy while preserving its sign, ATTNPO enables finer-grained credit assignment without introducing conflicts with the outcome-level supervision. Notably, advantage scaling is applied only to correct responses, as identifying essential versus redundant steps in incorrect responses is inherently noisy and can lead to spurious credit assignment.

In this work, we follow the TLMRE (Arora and Zanette, 2025) setup to compute the outcome-level advantage. Specifically, the reward is defined as

$$r_i = \mathbb{I}[o_i \text{ is correct}] (1 - \alpha \cdot \sigma(f(o_i))), \quad (5)$$

$$f(o_i) = \sigma((\text{len}(o_i) - \text{mean}(q)) / \text{std}(q)). \quad (6)$$

Here, $|o_i|$ is the length of o_i , $\text{mean}(q)$ and $\text{std}(q)$ denote the mean and standard deviation of the

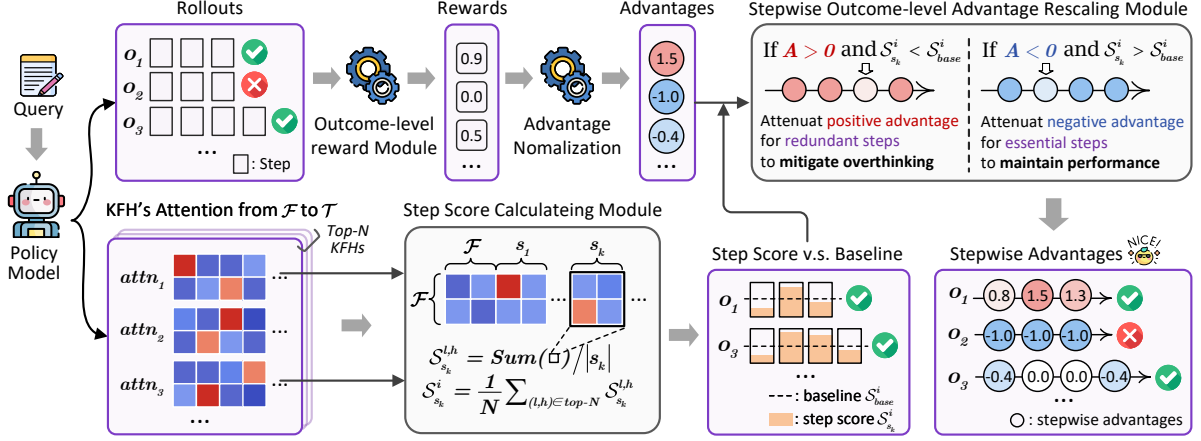


Figure 3: The overall framework of ATTNPO.

lengths of correct rollouts sampled by q . α is a hyperparameter. Following them, we use the RLOO advantage estimator: $A^i = r_i - \frac{1}{G-1} \sum_{j \neq i} r_j$.

4.2 Pos-Adv Attenuation for Redundant-Step

When a response has a positive outcome-level advantage A^i (Pos-Adv), its generation probability is reinforced. Thus, we attenuate advantages for relatively redundant steps to avoid over-encouraging them, thereby mitigating overthinking.

First, for each step s_k in response o_i , we compute a redundancy-related score $S_{s_k}^i$ by averaging the attention-based step scores (Eq. 1) over the top- N KFHS ranked by SRA scores, i.e., $S_{s_k}^i = \frac{1}{k} \sum_{(l,h) \in \text{top-}N} S_{s_k}^{l,h}$.

Then, we compute a response-specific baseline score S_{base}^i , which serves as a threshold for distinguishing redundant and essential steps. Specifically, steps with $S_{s_k}^i < S_{\text{base}}^i$ are treated as redundant, while the remaining steps are considered essential. The baseline score S_{base}^i is defined as:

$$S_{\text{base}}^i \triangleq p_i^\beta \cdot \frac{|\mathcal{F}_i|}{|o_i|}, \text{ where } p_i = \frac{1}{G} \sum_{j=1}^G \mathbb{I}[o_j \text{ is correct}]. \quad (7)$$

Here, $|\mathcal{F}_i|/|o_i|$ measures the average attention received by response tokens from the final solution \mathcal{F}_i , and thus provides a natural reference for identifying relatively under-attended steps as redundant. The term p_i serves as a proxy for problem difficulty, estimated by the average rollout accuracy, and β is a hyperparameter controlling its influence. The baseline is difficulty-aware, as harder problems demand stricter redundancy criteria due to the increased importance of correctness.

Finally, for correct responses with positive advantages, we define the stepwise scaling factor $\gamma_{s_k}^i$

in a difficulty-aware and scheduled manner:

$$\gamma_{s_k}^i = \begin{cases} (1 - \delta) \cdot p_i^\lambda \cdot \left(\frac{S_{s_k}^i}{S_{\text{base}}^i} \right) + \delta, & \text{if } S_{s_k}^i < S_{\text{base}}^i \\ & \wedge t > T \cdot p_i, \\ 1, & \text{otherwise.} \end{cases} \quad (8)$$

Here, the ratio $S_{s_k}^i/S_{\text{base}}^i$ serves as a measure of step-level redundancy, and advantage attenuation is applied only to potentially redundant steps ($S_{s_k}^i < S_{\text{base}}^i$). The factor p_i^λ controls the attenuation strength with a difficulty-aware magnitude, weakening attenuation for harder problems, where λ is a hyperparameter. t and T denote the current and total training steps, respectively, and the condition $t > T \cdot p_i$ defines a difficulty-aware schedule that delays advantage attenuation on harder problems, by allowing adequate exploration in early stages. Finally, δ specifies a lower bound on $\gamma_{s_k}^i$.

4.3 Neg-Adv Attenuation for Essential-Step

When a correct response receives a negative outcome-level advantage A^i (Neg-Adv), its generation probability is suppressed. In this case, we attenuate the penalties on essential steps; as a result, the remaining penalty is relatively focused on redundant ones, thereby avoiding performance degradation caused by penalizing essential reasoning. Specifically, for correct responses with negative advantages, we define the $\gamma_{s_k}^i$ as follows:

$$\gamma_{s_k}^i = \begin{cases} 0, & \text{if } S_{s_k}^i > S_{\text{base}}^i, \\ 1, & \text{otherwise.} \end{cases} \quad (9)$$

Here, we clip negative advantages of potentially essential steps ($S_{s_k}^i > S_{\text{base}}^i$) of correct responses to zero to avoid over-penalization on them.

Methods	GSM8K			MATH500			AMC2023			Olympiad			AIME2024			AIME2025			Macro Avg		
	Acc.	Tok.	AES	Acc.	Tok.	AES	Acc.	Tok.	AES	Acc.	Tok.	AES	Acc.	Tok.	AES	Acc.	Tok.	AES	Acc.	Tok.	AES
1.5B Model																					
DS-1.5B	78.8	1085	0	82.1	5534	0	65.9	7893	0	49.3	9034	0	28.1	12339	0	22.8	12143	0	54.5	8005	0
♥ AutoThink	83.0	568	0.64	84.0	2195	0.67	67.0	5059	0.41	51.6	5053	0.58	34.6	9514	0.92	21.8	7944	0.13	57.0	5056	0.56
♥ AdaptThink	83.1	<u>480</u>	0.72	82.0	1782	0.67	67.0	3510	0.61	50.8	3708	0.68	31.0	6670	0.77	23.0	7255	0.43	56.1	3901	0.65
♠ ThinkPrune	83.0	609	0.6	83.2	1938	0.69	73.2	3039	0.95	51.4	3535	0.74	27.1	<u>5631</u>	0.37	22.5	<u>5167</u>	0.51	56.7	<u>3320</u>	0.64
♠ DIET	-	-	-	83.0	3061	0.48	65.4	6425	0.15	-	-	-	31.8	10578	0.54	-	-	-	-	-	-
♠ ACPO	81.3	572	0.57	81.0	<u>1679</u>	0.63	-	-	-	-	-	-	30.0	6670	0.66	-	-	-	-	-	-
♠ Laser-D	83.4	863	0.38	84.2	1872	0.74	75.3	<u>2981</u>	1.05	54.4	4700	0.79	34.2	5750	1.19	23.1	6928	0.47	59.1	3849	0.77
♠ Laser-DE	80.4	820	0.31	83.5	1949	0.7	73.3	3080	0.95	54.4	5151	0.74	35.0	5789	1.27	24.2	7323	0.58	58.5	4019	0.76
♣ LC-R1	82.7	841	0.37	82.5	2233	0.61	61.7	3947	0.18	48.1	4546	0.38	23.6	7122	-0.37	21.2	6434	0.12	53.3	4187	0.22
♣ VSRM-R++	-	-	-	81.7	2597	0.51	64.7	4119	0.39	54.8	4388	0.85	29.5	6958	0.59	22.9	6892	0.45	-	-	-
♣ DEPO	-	-	-	87.2	2762	0.69	74.2	4388	0.82	-	-	-	30.8	7732	0.66	24.8	7649	0.63	-	-	-
♣ DECS	-	-	-	84.4	1817	0.76	75.4	2988	1.05	56.1	<u>3396</u>	<u>1.04</u>	31.2	5550	0.89	23.8	4965	0.72	-	-	-
♣ TLMRE	87.2	604	<u>0.76</u>	85.8	1915	<u>0.79</u>	79.1	3349	<u>1.18</u>	<u>56.9</u>	4270	0.99	<u>37.3</u>	7024	<u>1.41</u>	26.6	7117	0.91	62.2	4046	<u>1.01</u>
♣ ATTNPO	<u>87.0</u>	393	0.95	<u>86.0</u>	1318	0.9	<u>77.8</u>	2638	1.21	57.1	3133	1.13	37.7	5713	1.56	<u>25.4</u>	5488	<u>0.89</u>	<u>61.8</u>	3114	1.11
7B Model																					
DS-7B	88.2	639	0	92.0	3593	0	87.1	5977	0	65.1	7358	0	52.9	10490	0	35.8	11307	0	70.2	6561	0
♥ AutoThink	91.1	866	-0.26	91.2	2146	0.36	83.3	4645	0	65.5	5133	0.32	54.8	8051	0.34	36.2	8608	0.27	70.4	4908	0.17
♥ AdaptThink	91.0	309	0.61	92.0	1875	0.48	87.5	3287	0.46	65.5	5574	0.26	55.6	8599	0.33	36.2	9523	0.19	71.3	4861	0.39
♠ DIET	-	-	-	92.1	3187	0.12	82.6	6075	-0.27	-	-	-	57.9	10124	0.32	-	-	-	-	-	-
♠ ACPO	88.3	<u>413</u>	0.36	91.6	<u>1405</u>	<u>0.59</u>	-	-	-	-	-	-	52.8	4520	0.56	-	-	-	-	-	-
♠ Laser-D	87.3	804	-0.31	92.2	1836	0.5	<u>90.0</u>	2694	<u>0.65</u>	66.7	3914	0.54	<u>58.3</u>	5379	0.79	38.0	6167	0.64	72.1	3466	0.47
♠ Laser-DE	82.2	789	-0.57	92.0	1658	0.54	89.1	<u>2612</u>	0.63	66.8	3643	0.58	55.8	<u>4969</u>	0.69	37.5	6077	0.61	70.6	<u>3291</u>	0.41
♣ LC-R1	88.1	450	0.29	90.4	1568	0.48	79.1	3453	-0.04	64.1	4144	0.36	23.6	6904	-2.42	36.2	7150	0.4	63.6	3945	-0.16
♣ S-GRPO	93.8	906	-0.23	92.4	2252	0.39	87.5	3494	0.43	<u>69.7</u>	3914	0.68	56.0	7377	0.47	36.0	7908	0.32	72.6	4308	0.34
♣ VSRM-R++	-	-	-	89.8	2044	0.31	80.9	3704	0.02	66.1	5470	0.3	52.2	6773	0.29	36.4	6953	0.44	-	-	-
♣ DEPO	-	-	-	94.4	2318	0.43	90.5	3215	0.58	-	-	-	52.7	6580	0.35	39.2	7092	<u>0.66</u>	-	-	-
♣ DECS	-	-	-	93.0	1728	0.55	89.0	2772	0.6	70.3	<u>3283</u>	0.79	51.3	5277	0.35	36.4	<u>5516</u>	<u>0.56</u>	-	-	-
♣ TLMRE	91.1	515	0.29	<u>93.8</u>	1700	0.59	89.9	2967	0.6	69.6	4454	0.6	58.5	7462	0.61	39.2	8301	0.55	73.7	4233	<u>0.54</u>
♣ ATTNPO	<u>92.4</u>	446	<u>0.44</u>	92.8	1300	0.66	89.6	2303	0.7	68.7	2912	<u>0.77</u>	57.2	5214	<u>0.75</u>	<u>38.1</u>	5359	0.72	<u>73.1</u>	2922	0.67

Table 1: Acc. and Tok. across six benchmarks. Best scores are in **bold**, second-best are underlined. ♥, ♠, and ♣ denote adaptive-mode methods, outcome-supervised RL methods and process-supervised RL methods respectively.

5 Experiments

5.1 Experiment Setups.

Evaluation We evaluate our method on GSM8K (Cobbe et al., 2021), MATH500 (Lightman et al., 2024), AMC23, OlympiadBench (He et al., 2024), AIME2024, and AIME2025 as in-domain benchmarks, and on LiveCodeBenchv6 (Jain et al., 2025), GPQA-Diamond (Rein et al., 2024) and MMLU (Hendrycks et al., 2020) as held-out benchmarks, covering math, coding, and science tasks of varying difficulty. We sample 128 times for AIME2024 and AIME2025, and 4 times for all other benchmarks. We report Pass@1 (Acc.), response length (Tok.), and compute the Average Efficiency Score (AES) (Luo et al., 2025a) to provide a comprehensive assessment of both effectiveness and efficiency. Evaluation details are provided in Appendix B.2.

Training We train DeepSeek-R1-Distill-Qwen-1.5B (DS-1.5B) and DeepSeek-R1-Distill-Qwen-7B (DS-7B) on the DeepScaleR-Preview dataset (Luo et al., 2025b), using the VERL framework (Sheng et al., 2025). We set $\beta = 2$ (Eq. 7),

$\lambda = 2$ (Eq. 8), $\delta = 0$ for DS-1.5B and $\delta = 0.5$ for DS-7B, and fix $N = 3$ KFHs. Additional hyperparameters are reported in Appendix B.1.

Baselines We compare against three categories of methods for efficient reasoning: (1) Outcome-supervised RL methods, including TLMRE (Arora and Zanette, 2025), ThinkPrune (Hou et al., 2025), DIET (Chen et al., 2025a), ACPO (Cheng et al., 2025a), and Laser (Liu et al., 2025a); (2) Process-supervised RL methods, including LC-R1 (Cheng et al., 2025b), VSRM-R++ (Yue et al., 2025), S-GRPO (Dai et al., 2025), DEPO (Tan et al., 2025b), and DECS (Jiang et al., 2025); and (3) Adaptive-mode methods, including AdaptThink (Zhang et al., 2025a) and AutoThink (Tu et al., 2025), which are reported for reference. For TLMRE, we re-implement under the same experimental settings as ATTNPO to ensure fair comparison. For other methods, we either evaluate their publicly released models or report results from the original papers.

5.2 Main Results

ATTNPO consistently achieves the best or second-best AES across all six math benchmarks and at-

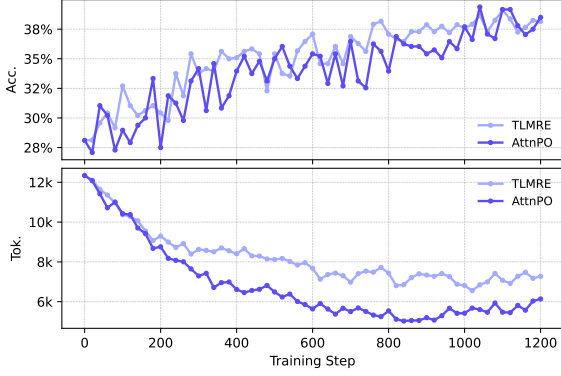


Figure 4: Training dynamics of TLMRE and ATTNPO on 1.5B scale on AIME2024 with 16 sampling runs.

tains the highest average AES across all datasets, indicating a superior efficiency–performance trade-off. As shown in Table 1, ATTNPO reduces the average length by 61% on the 1.5B model while improving pass@1 accuracy by +7.3 points. For example, on the challenging AIME2024 benchmark, ATTNPO achieves a +9.6 point accuracy gain with a 54% reduction in length. On the 7B model, it cuts tokens by 55% and achieves a +2.9 point accuracy gain, demonstrating efficiency and performance improvements across model scales. Notably, when compared with TLMRE, the outcome-level advantage used by ATTNPO, ATTNPO achieves much larger reductions in length (23% ↓ on 1.5B and 31% ↓ on 7B) with negligible loss in accuracy (0.4 points ↓ on 1.5B and 0.6 points ↓ 7B). The training dynamics in Figure 4 on AIME2024 show that ATTNPO initially underperforms TLMRE on harder tasks due to rapid length compression, but gradually improves and ultimately matches TLMRE in accuracy, indicating that its stepwise advantages enable continued performance gains despite aggressive compression.

5.3 Ablation Studies

In this section, we perform ablation studies on 1.5B models to assess different strategies and hyperparameters, with results shown in Figure 5.

Ablation for two Advantage Rescaling Components. Starting from the TLMRE, we sequentially add Positive-Advantage Attenuation (PA) and Negative-Advantage Attenuation (NA). Applying PA alone substantially reduces reasoning length, indicating that merely weakening the positive advantage of redundant steps provides sufficient optimization pressure to suppress unnecessary reasoning. Further introducing NA leaves the length almost unchanged while improving accuracy, sug-

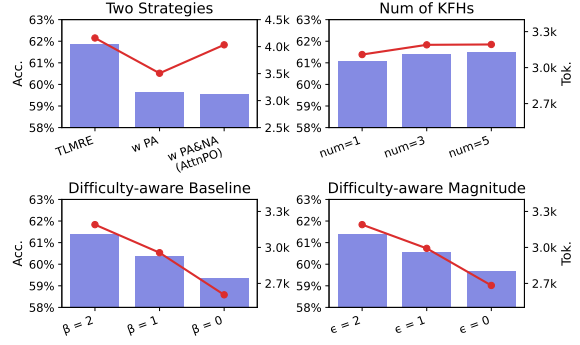


Figure 5: Ablation of strategies & hyperparameters; purple bars show Tok., red line shows Acc..

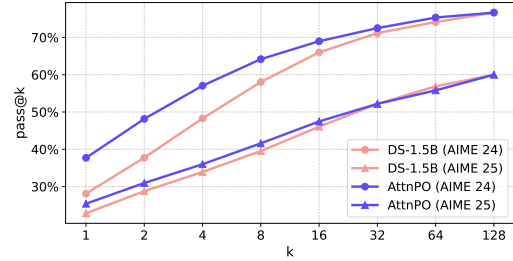


Figure 6: Change of pass@k with different k.

gesting that alleviating excessive penalties on necessary reasoning benefits reasoning performance.

Ablation for Num of KFHS. Increasing KFHS yields only marginal gains with rapidly diminishing returns, consistent with results in Section 3.2, suggesting that a small top-N set of KFHS provides sufficient learning signals.

Ablation for Difficulty-Aware Baselines and Magnitude. We ablate the difficulty-aware baseline β (Eq. 7) and magnitude λ (Eq. 8) to study their role in the performance–efficiency trade-off. Weakening (β or $\lambda = 1$) and removing (β or $\lambda = 0$) the difficulty-based modulation further shortens responses but causes increasing accuracy degradation, indicating that conservative compression on difficult problems is crucial for reducing verbosity while improving performance. Additionally, we provide an ablation study of the difficulty-aware schedule in Appendix C.3.

5.4 More Analyses

To better understand the effects of ATTNPO, we conduct further analysis on 1.5B models, unless otherwise specified.

Analyses of Exploration Capability. To assess whether ATTNPO’s substantial reduction in reasoning length compromises exploration, we evaluate pass@k ($k \in 2, 4, 8, 16, 32, 64, 128$) using 128

Model	LiveCodeBench			GPQA			MMLU		
	Acc.	Tok.	AES	Acc.	Tok.	AES	Acc.	Tok.	AES
1.5B Model									
DS-1.5B	25.3	10809	0.00	33.3	8304	0.00	41.7	1872	0.00
TLMRE	28.6	9077	0.55	39.2	5880	0.77	48.4	1228	0.83
ATTNPO	30.4	7567	0.90	36.4	3358	0.87	50.6	603	1.32
7B Model									
DS-7B	49.6	9095	0.00	44.6	6891	0.00	69.7	1245	0.00
TLMRE	53.5	3988	0.80	49.1	3988	0.72	70.0	712	0.44
ATTNPO	52.3	2873	0.85	48.6	2873	0.85	69.8	547	0.56

Table 2: OOD results on LiveCodeBench, GPQA, and MMLU. Best results are shown in **bold**.

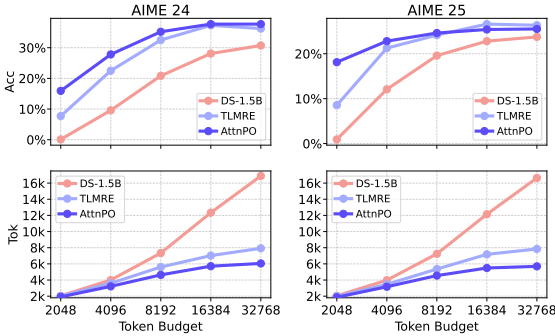


Figure 7: Acc. and Tok. under different token budgets.

samples on AIME2024 and AIME2025. As shown in Figure 6, in nearly all sampling budgets, ATTNPO match or outperform the original model. These results indicate that ATTNPO achieves significant efficiency gains while retaining much of the model’s exploratory capacity.

Analyses of Cross-Domain Capability. We analyze the model’s generalization ability on out-of-domain (OOD) tasks. In Table 2, ATTNPO substantially reduces reasoning length on both models, while maintaining or slightly improving accuracy. These results indicate that, despite being trained only on math, ATTNPO generalizes well to code and science tasks.

Analyses of Different Token Budget. We evaluate on AIME2024 and AIME2025 using five increasing token budgets (2048, 4096, 8192, 16384, 32768) to examine whether ATTNPO can improve performance under tight token constraints without sacrificing performance at higher limits. The results show that ATTNPO significantly outperforms both the base model and TLMRE under tight budgets, and remains superior to the base model while matching TLMRE at higher budgets with substantially shorter outputs, demonstrating the superiority of ATTNPO across a range of generation length constraints.

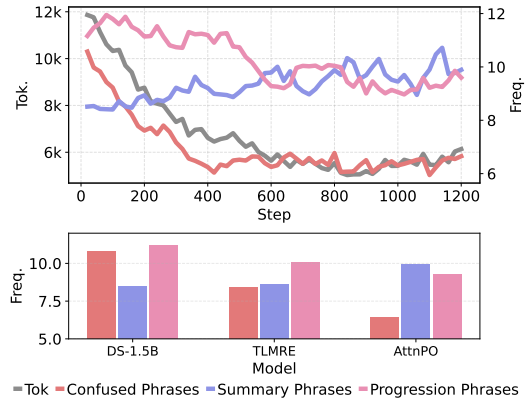


Figure 8: Changes in special phrases across training steps (top) and models (bottom). Freq. indicates the frequency of special phrases per 1k tokens.

Analyses of Model Behaviors. To better understand the response length reduction of ATTNPO, we analyze the frequency of special phrases with different reasoning behaviors (introduced in section 2) on AIME2024. We show the evolution of these special phrases during ATTNPO training (Figure 8, top) and the final behavioral differences among ATTNPO, TLMRE, and DS-1.5B (Figure 8, bottom). The frequency of Confused Phrases drops sharply during ATTNPO training, indicating that ATTNPO effectively reduces redundant self-reflection, this trend is also observed in cross-model comparisons. The slight increase in Summary Phrases suggests growing confidence in intermediate reasoning results. Meanwhile, the mild decrease in Progression Phrases implies more consolidated reasoning by skipping trivial intermediate steps.

6 Related Work

We first introduce overthinking in LRMs and common mitigation approaches, then focus on outcome-supervised and process-supervised RL methods which most closely related to our work.

Overthinking and Its Mitigation. Benefiting from the RLVR training paradigm (Tan et al., 2025a; Li et al., 2026; Yu et al., 2026; Yang et al., 2026), LRMs have achieved strong performance on complex reasoning tasks (An et al., 2025). However, they often exhibit overthinking by generating redundant reasoning (Wei et al., 2025; Dong et al., 2026), which can even degrade performance on certain tasks (Yuan and Zhang, 2025). Mitigating overthinking remains a challenging problem. Existing methods mainly fall into five categories: (1) RL

with redundancy-based reward design, which we discuss in detail below; (2) curating concise CoT data for SFT or DPO (Yuan et al., 2026; Ma et al., 2025); (3) compressing reasoning steps into fewer latent representations (Deng et al., 2025); and (4) dynamic reasoning strategy selection (Wang et al., 2026; Zhang et al., 2025a); (5) inference-time intervention methods (Liu et al., 2025b; Renze and Guven, 2024; Ling et al., 2026; Zhu et al., 2026).

Outcome-supervised RL method. This line of work typically adopts response-level rewards combining correctness with length-aware penalties. Some approaches explicitly subtract length-correlated penalty terms (Arora and Zanette, 2025; Aggarwal and Welleck, 2025; Team et al., 2025; Yi et al., 2025), while others further refine this by jointly modeling response length and question difficulty (Chen et al., 2025a; Liu et al., 2025a; Xiang et al., 2025). However, these methods treat the output as a whole and fail to distinguish efficient from inefficient reasoning segments.

Process-supervised RL method. Some methods introduce process supervision to intermediate reasoning steps to distinguish necessary from redundant reasoning. One line of work incorporates process supervision through additional sampling. S-GRPO (Dai et al., 2025), VSRM (Yue et al., 2025), SHAPE (Ai et al., 2026) and SPAE (Wu et al., 2026) achieve fine-grained credit assignment by allowing early termination at intermediate steps, with the model directly samples the final solution. And ROSE (Zhao et al., 2026) extends process supervision to tree-structured rollouts with concise reasoning incentives. Another line of work introduces auxiliary reward models (Liu et al., 2026). DECS (Jiang et al., 2025) and DEPO (Tan et al., 2025b) train an auxiliary to identify when the correct answer first appears and penalize subsequent redundant tokens. Additionally, since tokens with different entropy may play different roles in reasoning (Wang et al., 2025; Zhang et al., 2026), some work performs entropy- or confidence-aware advantage shaping (Jin et al., 2025; Liu et al., 2025c; Xie et al., 2025). However, these methods still rely on surface-level entropy signals and often lead to longer reasoning trajectories.

7 Conclusion

In this paper, we introduce ATTNPO, a low-overhead process-supervised RL framework that

leverages KFHs’ attention signals for stepwise credit assignment. By identifying KFHs that focus on essential steps and designing targeted sub-strategies to suppress redundancy while preserving necessary reasoning, ATTNPO effectively mitigates overthinking without sacrificing performance. These results highlight KFH-guided process supervision as an efficient way to improve reasoning efficiency.

Limitations

We discuss several limitations of our work in this section. (1) Owing to limited computational resources, our experiments are conducted only on 1.5B and 7B models. Nonetheless, the results consistently demonstrate the effectiveness of ATTNPO across different model scales. (2) In line with most prior open-sourced studies, we train our models exclusively on mathematical datasets, as they are readily available and provide accurate, verifiable reward signals. Although evaluations on LiveCodeBench, GPQA, and MMLU indicate that ATTNPO generalizes well to OOD scenarios, we anticipate further improvements with the availability of more verifiable training data from broader domains.

Acknowledgments

We would like to thank the anonymous reviewers, the meta-reviewer, as well as the area chairs and program chairs for their valuable comments and efforts. This work is supported by the National Natural Science Foundation of China (Grant No. 62572465).

References

- Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*.
- Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. [Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267, Bangkok, Thailand. Association for Computational Linguistics.
- Zhengyang Ai, Zikang Shan, Xiaodong Ai, Jingxian Tang, Hangkai Hu, and Pinyan Lu. 2026. Shape: Stage-aware hierarchical advantage via potential estimation for llm reasoning. *arXiv preprint arXiv:2604.06636*.

- AI-MO. 2024. Amc 2023 dataset.
- Shengnan An, Xunliang Cai, Xuezhi Cao, Xiaoyu Li, Yehao Lin, Junlin Liu, Xinxuan Lv, Dan Ma, Xuanlin Wang, Ziwen Wang, and 1 others. 2025. Amo-bench: Large language models still struggle in high school math competitions. *arXiv preprint arXiv:2510.26768*.
- Daman Arora and Andrea Zanette. 2025. **Training language models to reason efficiently**. In *The Thirtieth Annual Conference on Neural Information Processing Systems*.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. **Evaluating large language models trained on code**. *Preprint*, arXiv:2107.03374.
- Weize Chen, Jiarui Yuan, Tailin Jin, Ning Ding, Huimin Chen, Zhiyuan Liu, and Maosong Sun. 2025a. The overthinker’s diet: Cutting token calories with difficulty-aware training. *NeurIPS*.
- Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025b. **Do not think that much for 2+3=? on the overthinking of o1-like llms**. *Preprint*, arXiv:2412.21187.
- Yao Chen, Yilong Chen, Yinqi Yang, Junyuan Shang, Zhenyu Zhang, Zefeng Zhang, Shuaiyi Nie, Shuo-huan Wang, Yu Sun, Hua Wu, and 1 others. 2026. Sparse growing transformer: Training-time sparse depth allocation via progressive attention looping. *arXiv preprint arXiv:2603.23998*.
- Xiaoxue Cheng, Junyi Li, Zhenduo Zhang, Xinyu Tang, Wayne Xin Zhao, Xinyu Kong, and Zhiqiang Zhang. 2025a. Incentivizing dual process thinking for efficient large language model reasoning. *arXiv preprint arXiv:2505.16315*.
- Zhengxiang Cheng, Dongping Chen, Mingyang Fu, and Tianyi Zhou. 2025b. Optimizing length compression in large reasoning models. *arXiv preprint arXiv:2506.14755*.
- Daewon Choi, Jimin Lee, Jihoon Tack, Woomin Song, Saket Dingliwal, Sai Muralidhar Jayanthi, Bhavana Ganesh, Jinwoo Shin, Aram Galstyan, and Sravan Babu Bodapati. 2025. **Think clearly: Improving reasoning via redundant token pruning**. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21437–21451, Suzhou, China. Association for Computational Linguistics.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. **Training verifiers to solve math word problems**. *Preprint*, arXiv:2110.14168.
- Muzhi Dai, Chenxu Yang, and Qingyi Si. 2025. S-grpo: Early exit via reinforcement learning in reasoning models. *NeurIPS*.
- Jingcheng Deng, Liang Pang, Zihao Wei, Shichen Xu, Zenghao Duan, Kun Xu, Yang Song, Huawei Shen, and Xueqi Cheng. 2025. Latent reasoning in llms as a vocabulary-space superposition. *arXiv preprint arXiv:2510.15522*.
- Haonan Dong, Kehan Jiang, Haoran Ye, Wenhao Zhu, Zhaolu Kang, and Guojie Song. 2026. **Neureasoner: Towards explainable, controllable, and unified reasoning via mixture-of-neurons**. *Preprint*, arXiv:2604.02972.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. **DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning**. *Nature*, 645(8081):633–638.
- Hasan Abed Al Kader Hammoud, Hani Itani, and Bernard Ghanem. 2025. **Beyond the last answer: Your reasoning trace uncovers more than you think**. *Preprint*, arXiv:2504.20708.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 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, Bangkok, Thailand. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Bairu Hou, Yang Zhang, Jiabao Ji, Yujian Liu, Kaizhi Qian, Jacob Andreas, and Shiyu Chang. 2025. **Thinkprune: Pruning long chain-of-thought of llms via reinforcement learning**. *arXiv preprint arXiv:2504.01296*.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. **Livecodebench: Holistic and contamination free evaluation of large language models for code**. In *The Thirtieth International Conference on Learning Representations*.
- Kehan Jiang, Haonan Dong, Zhaolu Kang, Zhengzhou Zhu, and Guojie Song. 2026. **Foe: Forest of errors makes the first solution the best in large reasoning models**. *arXiv preprint arXiv:2604.02967*.

- Shuyang Jiang, Yusheng Liao, Ya Zhang, Yanfeng Wang, and Yu Wang. 2025. Overthinking reduction with decoupled rewards and curriculum data scheduling. *arXiv preprint arXiv:2509.25827*.
- Renren Jin, Pengzhi Gao, Yuqi Ren, Zhuowen Han, Tongxuan Zhang, Wuwei Huang, Wei Liu, Jian Luan, and Deyi Xiong. 2025. Revisiting entropy in reinforcement learning for large reasoning models. *arXiv preprint arXiv:2511.05993*.
- Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Libo Qin, Ruihan Chen, Baohang Li, Kui Jiang, Yaowei Wang, Ting Liu, and 1 others. 2025a. Cai: Caption-sensitive attention intervention for mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2506.23590*.
- Ruoran Li, Xinghua Zhang, Haiyang Yu, Shitong Duan, Xiang Li, Wenxin Xiang, Chonghua Liao, Xudong Guo, Yongbin Li, and Jinli Suo. 2026. Mempo: Self-memory policy optimization for long-horizon agents. *arXiv preprint arXiv:2603.00680*.
- Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhiwei Li, Bao-Long Bi, Ling-Rui Mei, Junfeng Fang, Xiao Liang, Zhi-jiang Guo, and 2 others. 2025b. From system 1 to system 2: A survey of reasoning large language models. *Preprint*, arXiv:2502.17419.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*.
- Guoming Ling, Zhongzhan Huang, Yupei Lin, Junxin Li, Shanshan Zhong, Hefeng Wu, and Liang Lin. 2026. Neural chain-of-thought search: Searching the optimal reasoning path to enhance large language models. *arXiv preprint arXiv:2601.11340*.
- Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. 2026. Openrubrics: Towards scalable synthetic rubric generation for reward modeling and llm alignment. *Preprint*, arXiv:2510.07743.
- Wei Liu, Ruochen Zhou, Yiyun Deng, Yuzhen Huang, Junteng Liu, Yuntian Deng, Yizhe Zhang, and Junxian He. 2025a. Learn to reason efficiently with adaptive length-based reward shaping. *Preprint*, arXiv:2505.15612.
- Yule Liu, Jingyi Zheng, Zhen Sun, Zifan Peng, Wenhan Dong, Zeyang Sha, Shiwen Cui, Weiqiang Wang, and Xinlei He. 2025b. Thought manipulation: External thought can be efficient for large reasoning models. *arXiv preprint arXiv:2504.13626*.
- Zheng Liu, Mengjie Liu, Siwei Wen, Mengzhang Cai, Bin Cui, Conghui He, and Wentao Zhang. 2025c. From uniform to heterogeneous: Tailoring policy optimization to every token’s nature. *arXiv preprint arXiv:2509.16591*.
- Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. 2025a. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. In *2nd AI for Math Workshop @ ICML 2025*.
- Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. 2025b. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. <https://pretty-radio-b75.notion.site/DeepScaler-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2>. Notion Blog.
- Xinyin Ma, Guangnian Wan, Runpeng Yu, Gongfan Fang, and Xinchao Wang. 2025. Cot-valve: Length-compressible chain-of-thought tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6025–6035.
- Mathematical Association of America. 2025a. American invitational mathematics examination 2024.
- Mathematical Association of America. 2025b. American invitational mathematics examination 2025.
- OpenAI. 2024. Learning to reason with LLMs. <https://openai.com/index/learning-to-reason-with-llms/>.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.
- Matthew Renze and Erhan Guven. 2024. The benefits of a concise chain of thought on problem-solving in large language models. *arXiv preprint arXiv:2401.05618*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

- Zelin Tan, Hejia Geng, Xiaohang Yu, Mulei Zhang, Guancheng Wan, Yifan Zhou, Qiang He, Xiangyuan Xue, Heng Zhou, Yutao Fan, and 1 others. 2025a. Scaling behaviors of llm reinforcement learning post-training: An empirical study in mathematical reasoning. *arXiv preprint arXiv:2509.25300*.
- Zezhong Tan, Hang Gao, Xinhong Ma, Feng Zhang, and Ziqiang Dong. 2025b. Towards flash thinking via decoupled advantage policy optimization. *arXiv preprint arXiv:2510.15374*.
- Kimi Team, Angang Du, Bofei Gao, Bawei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, and 75 others. 2025. [Kimi k1.5: Scaling reinforcement learning with llms](#). *CoRR*, abs/2501.12599.
- Songjun Tu, Jiahao Lin, Qichao Zhang, Xiangyu Tian, Linjing Li, Xiangyuan Lan, and Dongbin Zhao. 2025. Learning when to think: Shaping adaptive reasoning in r1-style models via multi-stage rl. *arXiv preprint arXiv:2505.10832*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Minzheng Wang, Yongbin Li, Haobo Wang, Xinghua Zhang, Nan Xu, Bingli Wu, Fei Huang, Haiyang Yu, and Wenji Mao. 2026. Adaptive social learning via mode policy optimization for language agents. In *The Fourteenth International Conference on Learning Representations*.
- Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.
- Zihao Wei, Liang Pang, Jiahao Liu, Wenjie Shi, Jingcheng Deng, Shicheng Xu, Zenghao Duan, Fei Sun, Huawei Shen, and Xueqi Cheng. 2025. The evolution of thought: Tracking llm overthinking via reasoning dynamics analysis. *arXiv preprint arXiv:2508.17627*.
- Fei Wu, Zhenrong Zhang, Qikai Chang, Jianshu Zhang, Quan Liu, and Jun Du. 2026. Step potential advantage estimation: Harnessing intermediate confidence and correctness for efficient mathematical reasoning. *arXiv preprint arXiv:2601.03823*.
- Violet Xiang, Chase Blagden, Rafael Rafailov, Nathan Lile, Sang Truong, Chelsea Finn, and Nick Haber. 2025. Just enough thinking: Efficient reasoning with adaptive length penalties reinforcement learning. *arXiv preprint arXiv:2506.05256*.
- Can Xie, Ruotong Pan, Xiangyu Wu, Yunfei Zhang, Jiayi Fu, Tingting Gao, and Guorui Zhou. 2025. Unlocking exploration in rlvr: Uncertainty-aware advantage shaping for deeper reasoning. *arXiv preprint arXiv:2510.10649*.
- Lei Yang, Wei Bi, Chenxi Sun, Renren Jin, and Deyi Xiong. 2026. Soup: Token-level single-sample mix-policy reinforcement learning for large language models. *arXiv preprint arXiv:2601.21476*.
- Jingyang Yi, Jiazheng Wang, and Sida Li. 2025. Shorterbetter: Guiding reasoning models to find optimal inference length for efficient reasoning. *NeurIPS*.
- Linhao Yu, Tianmeng Yang, Siyu Ding, Renren Jin, Naibin Gu, Xiangzhao Hao, Shuaiyi Nie, Deyi Xiong, Weichong Yin, Yu Sun, and Hua Wu. 2026. [Knowrl: Boosting llm reasoning via reinforcement learning with minimal-sufficient knowledge guidance](#). *Preprint*, arXiv:2604.12627.
- Haohan Yuan and Haopeng Zhang. 2025. Understanding llm reasoning for abstractive summarization. *arXiv preprint arXiv:2512.03503*.
- Hongyuan Yuan, Xinran He, Run Shao, Bolei He, Xianwei Xue, Mengke Chen, Qitong Pan, Haiwei Wang, and Haifeng Li. 2026. Graph-based chain-of-thought pruning for reducing redundant reflections in reasoning llms. *arXiv preprint arXiv:2604.05643*.
- Chuhuai Yue, Chengqi Dong, Yinan Gao, Hang He, Jiajun Chai, Guojun Yin, and Wei Lin. 2025. Promoting efficient reasoning with verifiable stepwise reward. *arXiv preprint arXiv:2508.10293*.
- Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. 2025a. [Adaptthink: Reasoning models can learn when to think](#). *arXiv preprint arXiv:2505.13417*.
- Wenyuan Zhang, Shuaiyi Nie, Xinghua Zhang, Zefeng Zhang, and Tingwen Liu. 2025b. [S1-bench: A simple benchmark for evaluating system 1 thinking capability of large reasoning models](#). *Preprint*, arXiv:2504.10368.
- Wenyuan Zhang, Xinghua Zhang, Haiyang Yu, Shuaiyi Nie, Bingli Wu, Juwei Yue, Tingwen Liu, and Yongbin Li. 2026. [Expseek: Self-triggered experience seeking for web agents](#). *arXiv preprint arXiv:2601.08605*.
- Ziqi Zhao, Zhaochun Ren, Jiahong Zou, Liu Yang, Zhiwei Xu, Xuri Ge, Zhumin Chen, Xinyu Ma, Daiting Shi, Shuaiqiang Wang, and 1 others. 2026. Reinforced efficient reasoning via semantically diverse exploration. *arXiv preprint arXiv:2601.05053*.
- Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. [Attention heads of large language models: A survey](#). *Preprint*, arXiv:2409.03752.
- Wei Zhu, Jian Zhang, Lixing Yu, Kun Yue, and Zhiwen Tang. 2026. [Dissecting failure dynamics in large language model reasoning](#). *Preprint*, arXiv:2604.14528.

A Appendix

A.1 Step Segmentation Algorithm

The thinking process \mathcal{T} can be decomposed into a sequence of meaningful intermediate steps, whose boundaries are often signaled by specific special phrases. These phrases mark key moments in the reasoning process, such as pauses, reflections, shifts in direction, or transitions to new stages (Hammoud et al., 2025; Choi et al., 2025). Accordingly, we segment the thinking process using a predefined set of special phrases, categorized into **confusion phrases** (e.g., ‘Wait’), which indicate reflection, verification, or branching; **progression phrases** (e.g., ‘First’), which advance the ongoing reasoning; and **summary phrases** (e.g., ‘Therefore’), which consolidate preceding content. The full set is listed in Table 3.

In LRMs, a double newline ($\backslash\backslash\n$) typically signals a natural semantic transition (Chen et al., 2025b). To avoid over-segmentation, we treat a special phrases as a segmentation boundary only when it is preceded by a double newline, except for confusion phrases. Since confusion phrases often trigger redundant verification or reflection, we do not impose the newline constraint on segments starting with them, keeping such segments separate from the preceding context to facilitate redundancy detection and penalization.

After initial segmentation, we apply adaptive merging to short segments to avoid semantically weak fragments, with the merging direction determined by the category of the initiating special phrases. Segments introduced by confusion phrases are always merged with the right neighbor to preserve their separation from prior reasoning. Segments introduced by progression phrases are merged based on context: if its right neighbor is a confusion-initiated segment, it is merged with the left neighbor to preserve the independence of the confusion segment; otherwise, it is merged with the right neighbor, as progression phases typically advance the reasoning flow. Segments introduced by summary phrases are merged with the left neighbor, as they serve to consolidate preceding reasoning.

Throughout segmentation and merging, we keep segments starting with confusion phrases separate from the preceding context, as they often correspond to redundant verification or reflection. Isolating such segments facilitates the identification and penalization of redundant reasoning.

We provide the pseudocode in Algorithm 1.

Algorithm 1 Step Segmentation Algorithm

Input: A thinking process \mathcal{T} , 3 predefined special-phrase sets $\mathcal{W}_{\text{conf}}$, $\mathcal{W}_{\text{prog}}$, and $\mathcal{W}_{\text{summ}}$, and a minimum length threshold t for each step

Output: Segmented steps $[s_1, s_2, \dots, s_K]$

```
Step 1: Initial Segmentation
1:  $\mathcal{B} \leftarrow []$   $\triangleright$  boundary positions (in left-to-right order)
2:  $\mathcal{P} \leftarrow []$   $\triangleright$  segmentation phrase at each boundary
3:  $\mathcal{W} \leftarrow \mathcal{W}_{\text{conf}} \cup \mathcal{W}_{\text{prog}} \cup \mathcal{W}_{\text{summ}}$ 
4: sort phrases in  $\mathcal{W}$  by decreasing length and compile a regex matcher R
5: for all matches  $(p, w)$  found by R when scanning  $\mathcal{T}$  left-to-right do  $\triangleright p$ : start position of the match;  $w$ : matched phrase
6:   if  $w \in \mathcal{W}_{\text{prog}} \cup \mathcal{W}_{\text{summ}}$  and  $w$  is not preceded by  $\backslash\backslash\n$  then
7:     continue
8:   end if
9:   append  $p$  to  $\mathcal{B}$ ; append  $w$  to  $\mathcal{P}$ 
10: end for
11: if  $\mathcal{B}[0] \neq 0$ , prepend 0 to  $\mathcal{B}$  and NONE to  $\mathcal{P}$ 
12: if  $\mathcal{B}[|\mathcal{B}| - 1] \neq |\mathcal{T}|$ , append  $|\mathcal{T}|$  to  $\mathcal{B}$ 
13: for  $i = 1$  to  $|\mathcal{B}| - 1$  do
14:    $s_i \leftarrow \mathcal{T}[\mathcal{B}[i] : \mathcal{B}[i + 1]]$ 
15: end for
16:  $K \leftarrow |\mathcal{B}| - 1$   $\triangleright$  total number of initial segmented steps
Step 2: Adaptive Merging
17: function MERGERIGHT( $i$ )
18:    $s_i \leftarrow s_i \circ s_{i+1}$ ; delete  $s_{i+1}$ 
19:   delete  $\mathcal{P}[i + 1]$ ; update  $K = K - 1$ 
20: end function
21: function MERGELEFT( $i$ )
22:    $s_{i-1} \leftarrow s_{i-1} \circ s_i$ ; delete  $s_i$ 
23:   delete  $\mathcal{P}[i]$ ; update  $K = K - 1$ 
24: end function
25:  $i \leftarrow 0$ 
26: while  $i < K$  do
27:   if  $\text{LEN}(s_i) < t$  then
28:     if  $\mathcal{P}[i] \in \mathcal{W}_{\text{conf}}$  then  $\triangleright$  confusion
29:       if  $i + 1 < K$  then
30:         MERGERIGHT( $i$ )
31:         continue
32:       end if
33:     else if  $\mathcal{P}[i] \in \mathcal{W}_{\text{summ}}$  then  $\triangleright$  summary
34:       if  $i - 1 \geq 0$  then
35:         MERGELEFT( $i$ )
36:          $i \leftarrow i - 1$ 
37:         continue
38:       end if
39:     else  $\triangleright$  progression
40:       if  $i + 1 < K$  and  $\mathcal{P}[i + 1] \in \mathcal{W}_{\text{conf}}$  then
41:         if  $i - 1 \geq 0$  then
42:           MERGELEFT( $i$ )
43:            $i \leftarrow i - 1$ 
44:         continue
45:         end if
46:       else
47:         if  $i + 1 < K$  then
48:           MERGERIGHT( $i$ )
49:           continue
50:         end if
51:       end if
52:     end if
53:   end if
54:    $i \leftarrow i + 1$ 
55: end while
56: return  $[s_1, s_2, \dots, s_K]$ 
```

Confusion phrases:

“Wait”, “But”, “However”, “Hold on”, “Looking back”, “I don’t see any errors”, “Hmm, wait”, “Hmm, no”, “Hmm, but”, “Hmm, just”, “Hmm, let me double-check”, “Hmm, let me verify”, “Hmm, let me make sure”, “Hmm, let me check”, “No”, “Let me double-check”, “Let me just double-check”, “Let’s double-check”, “Let’s just double-check”, “Just to double-check”, “Let me verify”, “Let me just verify”, “Let’s verify”, “Let’s just verify”, “Just to verify”, “Let me confirm”, “Let me just confirm”, “Let’s confirm”, “Let’s just confirm”, “Just to confirm”, “Let me check”, “Let me just check”, “Let’s check”, “Let’s just check”, “Just to check”, “Let me recap”, “Let me just recap”, “Let’s recap”, “Let’s just recap”, “Just to recap”, “Let me make sure”, “Let me just make sure”, “Let’s make sure”, “Just to make sure”, “So, is”, “Is it”, “Is that right?”, “Is there”, “Right?”, “Alternatively”, “Another approach”, “Another way”, “Another idea”, “Another thought”, “I guess another way”, “I guess another approach”, “Let me just think if there”

Progression phrases:

“Let me”, “Let’s”, “I”, “We”, “Okay”, “Hmm”, “Now”, “Alright”, “First”, “First off”, “Second”, “Third”, “Starting”, “Then”, “Next”, “Finally”, “Similarly”, “Again”, “In this case”, “Because”, “Given that”, “The problem gives”, “To”, “Since”

Summary phrases:

“So”, “Therefore”, “Thus”, “Hence”, “Hmm, so”, “Okay, so”, “Putting it all together”

Table 3: Special phrases used for step segmentation.

A.2 Details of the Probing Experiment

A.2.1 Details of Probe Data Construction

In this subsection, we describe the complete process for constructing the probing dataset. We select 300 questions from Deepscaler (Luo et al., 2025b) that are not difficult for DeepSeek-R1-Distill-Qwen-1.5B, on which the model achieves an average accuracy of at least 0.875 over eight sampled responses. This restriction is important because, on more challenging questions, LRMs tend to engage in exploratory reasoning. Although such reasoning may appear redundant, it can still be necessary for reaching correct solutions, making it difficult to reliably distinguish redundant steps from essential ones and potentially confounding the identification of Key-Focus Heads. For each selected question, we choose the correct response with median length among the eight samples and segment its thinking process into steps following the procedure described in Section A.1. Each reasoning step is then annotated by gemini-2.5-pro, gpt-5-2025-08-07, claude-sonnet-4-20250514 with the prompt provide in Table 7, and only steps with unanimous agreement are retained to reduce annotation noise.

The resulting inter-annotator agreement, measured by Krippendorff’s α , is 0.64 for DeepSeek-R1-Distill-Qwen-7B and 0.59 for DeepSeek-R1-Distill-Qwen-1.5B, indicating moderate agreement. We further retain only trajectories containing both essential and redundant steps, so that each probe instance provides an internal essential–redundant contrast, enabling attention comparisons within the same context.

A.2.2 Details of Probe Results

Here, we present the SRA heatmap across the checkpoints at 0, 500, and 1000 steps in Figure 11, and list the Pearson correlation coefficients of SRA among these checkpoints in Table 4. Additionally, in section 3.2, *Top-k selection by SRA* refers to ranking all attention heads by their average SRA scores and selecting the top k heads with the highest scores. *Greedy SRA-maximizing selection* refers to iteratively selecting attention heads by evaluating all remaining candidates at each step and greedily adding the head that yields the largest increase in the overall SRA score.

Model	0 vs 500	0 vs 1000
DS-R1-1.5B	0.92	0.88
DS-R1-7B	0.93	0.89

Table 4: Pearson correlation coefficients of SRA between checkpoints at 0, 500, and 1000 steps.

B Experimental Details

B.1 Training Hyperparameters

For the output-level reward in Eq. 5, we set $\alpha = 0.2$ for the 1.5B model and $\alpha = 0.1$ for the 7B model. This choice is based on our empirical observation that using $\alpha = 0.2$ for the 7B model leads to overly aggressive length compression and severe performance degradation, even without applying our method. For the KFHS used in both models, we select heads 16-2, 20-9, and 23-2 for the 1.5B model, and heads 22-4, 22-7, and 22-22 for the 7B model. For the 1.5B model, we train using a single 8×H800 node, while for the 7B model, training is conducted on four 8×H800 nodes. We present the other hyperparameters adopted during training in Table 5.

Param	
max response length	16384
batch size	128
rollout top-p	1.0
rollout temperature	1.0
learning rate	2e-6
kl loss Coef	0
rollout number	16
ϵ_{low}	0.2
ϵ_{high}	0.2
steps	1200

Table 5: Training hyperparameters.

B.2 Evaluation

B.2.1 Descriptions of Benchmarks

1. **AIME2024, AIME2025** (Mathematical Association of America, 2025a,b). These two datasets consist of high-school Olympiad-level assessments from the American Invitational Mathematics Examination (AIME) held in 2024 and 2025, respectively. Each dataset contains 30 highly challenging problems spanning algebra, geometry, and number theory.
2. **AMC23** (AI-MO, 2024). This dataset is sourced from the American Mathematics Competitions (AMC) system in 2023 and includes 40 problems with mixed and hybrid question formats.
3. **OlympiadBench** (He et al., 2024). This benchmark comprises a comprehensive collection of mathematical Olympiad problems from multiple countries. We select only the English-language math subset and retain problems that require numerical answers, resulting in a total of 581 evaluation problems.
4. **MATH500** (Lightman et al., 2024). This dataset is an advanced mathematics evaluation benchmark curated by OpenAI, containing 500 problems expressed with formal mathematical notation.
5. **GPQA-Diamond** (Rein et al., 2024). This dataset is a curated subset of the GPQA (Graduate-Level Google-Proof Q&A) benchmark and consists of 198 challenging multiple-choice questions authored and verified by domain experts in biology, physics, and chemistry.
6. **LiveCodeBench** (Jain et al., 2025). This benchmark is designed to evaluate the live code generation capabilities of large language

models, emphasizing immediate correctness and practical programming skills. We use version v6 of the dataset, which contains 1,055 problems in total.

7. **MMLU** (Hendrycks et al., 2020). MMLU is a massive multitask benchmark of multiple-choice questions spanning 57 subjects, including elementary mathematics, U.S. history, computer science, and law. Achieving high accuracy requires extensive world knowledge and strong problem-solving ability. We sample 50 questions from each category for evaluation.

B.2.2 Evaluation Setups

We sample 128 times for AIME2024 and AIME2025, and 4 times for all other benchmarks. We set the evaluation context size to 16K and the temperature to 0.6, as suggested in DeepSeek’s model cards.

B.2.3 Computation of Metrics

AES The AES score (Luo et al., 2025a) is calculated by jointly considering the pass@1 performance and the average token cost of the tuned policy relative to the base policy:

$$\text{AES} = \frac{L_{\text{base}} - L}{L_{\text{base}}} + \begin{cases} 3 \cdot \frac{\text{pass@1} - \text{pass@1}_{\text{base}}}{\text{pass@1}_{\text{base}}}, & \text{if } \text{pass@1} \geq \text{pass@1}_{\text{base}} \\ -5 \cdot \frac{\text{pass@1}_{\text{base}} - \text{pass@1}}{\text{pass@1}_{\text{base}}}, & \text{if } \text{pass@1} < \text{pass@1}_{\text{base}} \end{cases} \quad (10)$$

This metric captures both the relative reduction in token usage and its effect on model performance: it penalizes approaches that reduce performance and rewards those that yield improvements over the baseline.

Pass@k The pass@k metric (Chen et al., 2021) is computed as follows:

$$\text{pass@k} = 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}} \quad (11)$$

where n denotes the total number of samples and c denotes the number of correct samples. When $k = 1$, this metric reduces to the average accuracy across the n samples.

C Additional Experimental Results

C.1 Ablation for two sub-strategies

We present the ablation results for the two sub-strategies in Tables 6.

Methods	GSM8K		MATH500		AMC2023		Olympiad		AIME2024		AIME2025		Avg	
	Acc.	Tok.	Acc.	Tok.	Acc.	Tok.	Acc.	Tok.	Acc.	Tok.	Acc.	Tok.	Acc.	Tok.
TMLER	87.2	604	85.8	1915	79.1	3349	56.9	4270	37.3	7024	26.6	7117	62.2	4047
w PA	85.9	380	84.4	1295	76.8	2558	54.2	3240	36.8	5860	25.0	5670	60.5	3167
w PA&NA (ATTNPO)	87.0	393	86.0	1318	77.8	2638	57.1	3133	37.7	5713	25.4	5488	61.8	3114

Table 6: Ablation results for two sub-strategies, PA and NA.

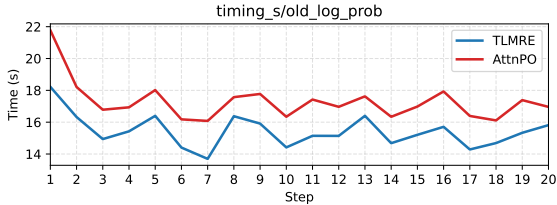


Figure 9: Time Comparison of ATTNPO and TLMRE on the 1.5B model.

C.2 Time Overhead Analysis of ATTNPO

We analyze the additional GPU time overhead introduced by ATTNPO, which mainly stems from computing attention-based scores for each reasoning step using KFH heads. Since our implementation is built on the VERL framework, this step is integrated into the computation of the log-probabilities of the rollout policy (referred to as `old_log_prob` in VERL). Specifically, during the model’s forward pass for computing importance sampling terms (e.g., log-probabilities and related quantities), we simultaneously compute the attention-based scores for each reasoning step, thereby avoiding additional forward passes and minimizing extra computational overhead. We report the training time of the 1.5B model under ATTNPO and TLMRE for the first 20 training steps as a comparison. We focus on the initial 20 steps because the generated response lengths under both methods are nearly identical in this stage, which effectively eliminates the confounding impact of length differences on the measured training time. The results in Figure 9 demonstrate that ATTNPO introduces almost no additional training time overhead.

C.3 Ablation for Difficulty-Aware Schedule

As shown in the training dynamics in Figure 10, introducing a difficulty-aware schedule results in a slower reduction of reasoning length during the early stages of training, but sustains a more consistent compression trend throughout training. In contrast, removing the difficulty-aware schedule leads to a faster initial decrease in reasoning length,

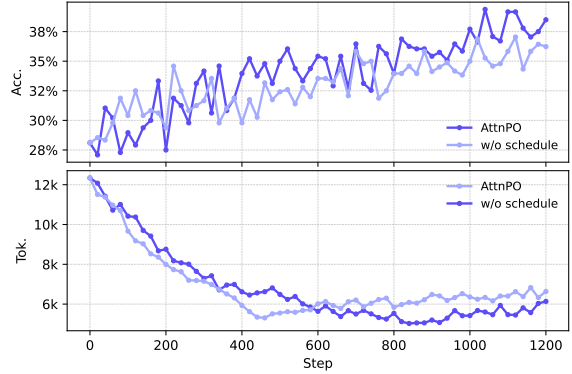


Figure 10: Ablation Results on the Difficulty-Aware Schedule on AIME2024 with 16 sampling runs.

followed by a clear plateau in later stages. This behavior suggests that delaying advantage attenuation on difficult problems allows the model to prioritize accuracy improvement in the early phase, preventing premature and excessive compression. Once performance is sufficiently improved, the model gains greater flexibility to further reduce reasoning length, leading to a more favorable performance–efficiency trade-off.

C.4 Case Study

We provide case studies of TLMRE and ATTNPO, with details presented in Table 12, 13, 14 and 15.

Prompts for Redundant and Necessary Step Annotation

Role & Task:

You are an expert evaluator of step-by-step reasoning quality for large language models.

I will give you a question and a corresponding model response (the response contains explicit step markers, e.g., <STA_2> means step 2 starts, <END_2> means step 2 ends). The final result in this response is correct. Please classify each step as: redundant / necessary / uncertain.

If a step contains only redundant content, label it as redundant; only necessary content, label it as necessary; if it contains both redundant and necessary content, label it as uncertain.

Category Definitions:

(1) **Redundant:** Steps that do not substantially advance the solution, or repeated/ineffective actions taken after the correct conclusion has already been reached, including but not limited to:

- **Redundant verification:** When the current reasoning is already correct, but the model performs an additional verification step. This is usually labeled as redundant. Only when the second verification is used to discover and correct an error, or to add essential justification, should it be labeled as necessary.
- **Redundant reflection:** When the current reasoning is already correct, but the model adds extra reflection or self-questioning that merely loops over the same reasoning without introducing new information, or even introduces incorrect information or baseless speculation. These are usually labeled as redundant. Only when such reflection is used to discover and correct an error, or to add essential justification, should it be labeled as necessary.
- **Redundant branches:** After arriving at the correct answer through valid reasoning, the model switches to another method to re-derive the correct answer. This is usually labeled as redundant.
- **Erroneous reasoning:** If a step contains incorrect statements and does not bring any effective progress, it should be considered redundant.

(2) **Necessary:** Steps that positively contribute to advancing the reasoning process, including but not limited to:

- **Normal computation/reasoning:** Necessary problem analysis, reasoning plans, formalization of the reasoning, intermediate calculations, fact retrieval, and explicit statements of conclusions that help push the reasoning forward should be labeled as necessary.
- **Necessary verification/reflection/branches:** When there are errors, contradictions, dead ends, or uncertainty due to multiple possible solutions, verification, reflection, or branching that is used to correct the direction, eliminate errors, or fill in missing conditions should be labeled as necessary.

(3) **Uncertain:**

- **Containing both redundant and necessary content:** If a single step includes both redundant content and necessary content, label it as uncertain.
- **Reasonable exploration:** If a step is not on the main line of reasoning but is relevant, correct, and introduces new information, label it as uncertain.
- **Borderline cases:** For borderline cases where you cannot clearly determine whether a step is redundant or necessary, label it as uncertain.

(4) **Borderline Cases:**

- Core planning used to guide subsequent reasoning steps is usually considered necessary.
- The model's final explicit statement of the final answer at the end of the reasoning process is generally considered necessary.

Question and Model Response:

Question:

{question}

Model response:

{response}

Output Requirements:

First, output the reasoning for your judgment on each step, in the following format:

Step index: reasoning for the judgment

Then output the classification result for each step, in the following format:

Step index: category (use numbers, 1 = necessary, 2 = redundant, 3 = uncertain; for example, 0:1 means step 0 is necessary)

Example output (strictly follow this format):

[Reasoning for each step]

0: Reasoning for the judgment...

1: Reasoning for the judgment...

2: Reasoning for the judgment...

[Classification results]

0:1

1:2

2:3

Begin your answer:

Table 7: Prompts for Redundant and Necessary Step Annotation.

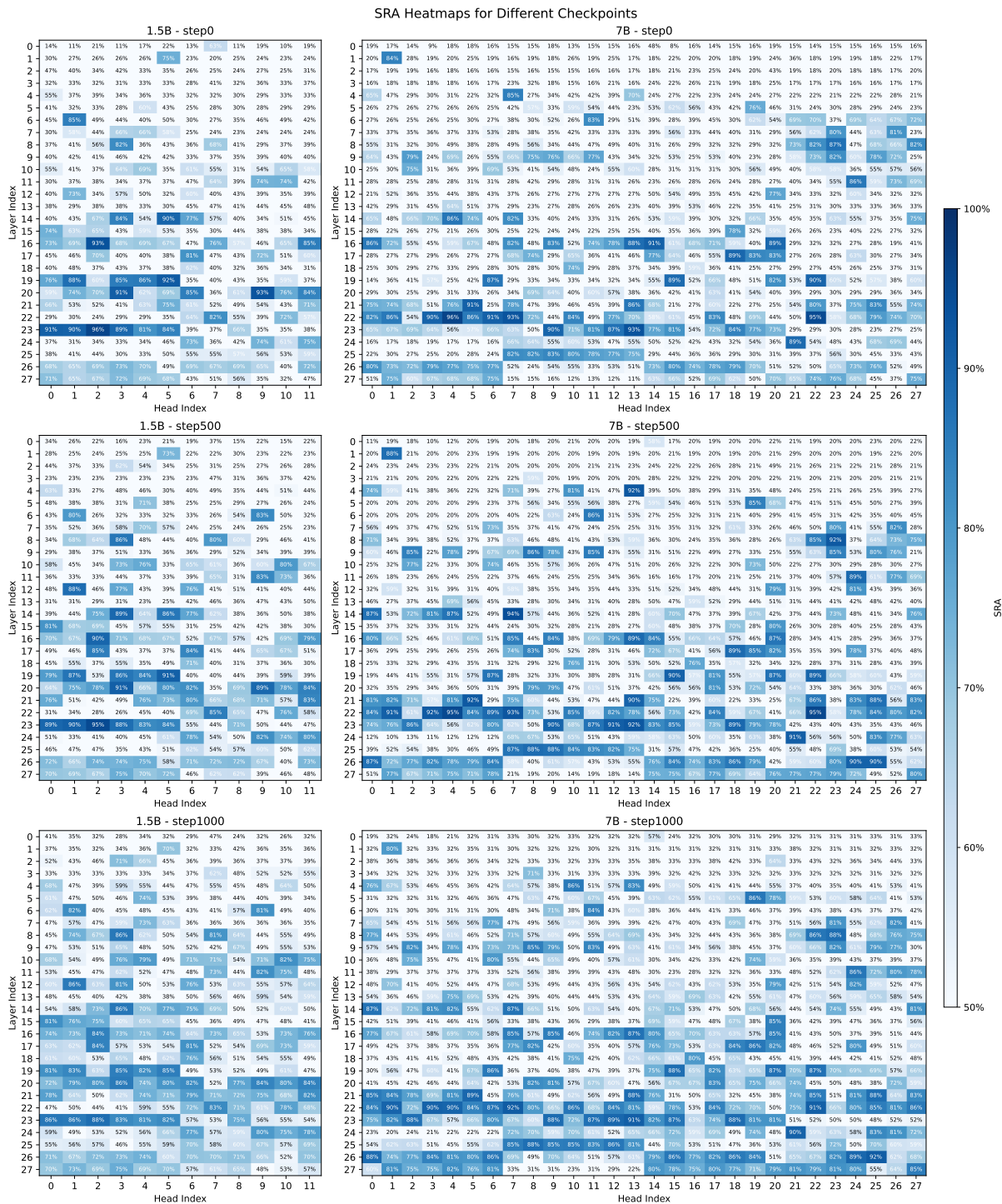


Figure 11: SRA heatmap for all models.

Question (MATH500): If $f(x) = \frac{3x-2}{x-2}$, what is the value of $f(-2) + f(-1) + f(0)$? Express your answer as a common fraction.

TLMRE: Okay, so I need to find the value of $f(-2) + f(-1) + f(0)$ where $f(x)$ is given by $(3x - 2)/(x - 2)$. Hmm, let me break this down step by step. First, let me compute $f(-2)$. I substitute $x = -2$ into the function: $f(-2) = (3 \cdot (-2) - 2)/(-2 - 2) = (-6 - 2)/(-4) = (-8)/(-4) = 2$. Alright, $f(-2)$ is 2. Next, $f(-1)$. Substitute $x = -1$: $f(-1) = (3 \cdot (-1) - 2)/(-1 - 2) = (-3 - 2)/(-3) = (-5)/(-3) = 5/3$. So, $f(-1)$ is $5/3$. Now, $f(0)$. Substitute $x = 0$: $f(0) = (3 \cdot 0 - 2)/(0 - 2) = (-2)/(-2) = 1$. So, $f(0)$ is 1. Now, add them all together: $f(-2) + f(-1) + f(0) = 2 + 5/3 + 1$. Convert all to thirds: 2 is $6/3$, 1 is $3/3$, so: $6/3 + 5/3 + 3/3 = (6 + 5 + 3)/3 = 14/3$. Wait, that seems correct? Let me double-check each calculation. $f(-2)$: $(3 \cdot (-2) - 2) = -6 - 2 = -8$; denominator $(-2 - 2) = -4$; $-8 / -4 = 2$. Correct. $f(-1)$: $(3 \cdot (-1) - 2) = -3 - 2 = -5$; denominator $(-1 - 2) = -3$; $-5 / -3 = 5/3$. Correct. $f(0)$: $(0 - 2)/(-2) = -2 / -2 = 1$. Correct. Sum: $2 + 5/3 + 1$. $2 + 1 = 3$, $3 + 5/3 = 3 + 1.666\dots = 4.666\dots$, which is $14/3$. Yeah, $14/3$ is correct as a common fraction. **Final Answer** The

value of $(f(-2) + f(-1) + f(0))$ is $\boxed{\frac{14}{3}}$.

ATTNPO-1.5B (ours):

Okay, so I have this function $f(x) = (3x - 2)/(x - 2)$, and I need to find $f(-2) + f(-1) + f(0)$. Hmm, let's take it step by step. First, let's find $f(-2)$. Plugging $x = -2$ into the function: $f(-2) = (3 \cdot (-2) - 2)/(-2 - 2) = (-6 - 2)/(-4) = (-8)/(-4) = 2$. Okay, that's straightforward. Next, $f(-1)$. Plugging $x = -1$: $f(-1) = (3 \cdot (-1) - 2)/(-1 - 2) = (-3 - 2)/(-3) = (-5)/(-3) = 5/3$. So that's $5/3$. Then, $f(0)$. Plugging $x = 0$: $f(0) = (3 \cdot 0 - 2)/(0 - 2) = (-2)/(-2) = 1$. That's simple too. Now, adding them up: $f(-2) + f(-1) + f(0) = 2 + 5/3 + 1$. Let's convert everything to thirds: $2 = 6/3$, $1 = 3/3$. So $6/3 + 5/3 + 3/3 = (6 + 5 + 3)/3 = 14/3$. **Final Answer** $\boxed{\frac{14}{3}}$.

Figure 12: This is a simple problem from MATH500, where an unnecessarily reflective reasoning process is applied by TLMRE-1.5B. In contrast, the AttnPO-1.5B provides a concise solution.

Question (AIME 2024): Every morning Aya goes for a 9-kilometer-long walk and stops at a coffee shop afterwards. When she walks at a constant speed of s kilometers per hour, the walk takes her 4 hours, including t minutes spent in the coffee shop. When she walks $s + 2$ kilometers per hour, the walk takes her 2 hours and 24 minutes, including t minutes spent in the coffee shop. Suppose Aya walks at $s + \frac{1}{2}$ kilometers per hour. Find the number of minutes the walk takes her, including the t minutes spent in the coffee shop.

TLMRE: Okay, let's try to figure out this problem step by step. Aya walks 9 kilometers every morning and stops at a coffee shop. The time she takes depends on her walking speed. We need to find how long her walk will take if she walks at a speed of $s + \frac{1}{2}$ km/h, including the time spent in the coffee shop. First, let's parse the information given. When she walks at speed s km/h, the total time taken, including the coffee shop time, is 4 hours. When her speed increases to $s + 2$ km/h, the total time becomes 2 hours and 24 minutes, which is $2 + \frac{24}{60} = 2.4$ hours. **We need to find the time it takes when she walks at $s + \frac{1}{2}$ km/h. Let me denote: Let t be the time spent in the coffee shop in minutes. Let d be the distance she walks before entering the coffee shop, which is 9 km minus the coffee shop time converted to distance. Wait, no. Actually, the total distance is 9 km, but the coffee shop time is t minutes, but the walking time is different. Wait, maybe I need to model this differently.** Let me denote: Let walking speed be s km/h. Let t be the time spent in the coffee shop in hours. So, when she walks at s km/h, the total time is walking time + $t = 4$ hours. Walking time is $9/s$ hours. So, $9/s + t = 4$. Similarly, when her speed is $s + 2$ km/h, the total time is 2 hours and 24 minutes, which is $2 + \frac{24}{60} = 2.4$ hours. So, the walking time is $9/(s + 2)$ hours, and the total time is that plus t , which equals 2.4 hours. Therefore, we have two equations: 1. $9/s + t = 4$. 2. $9/(s + 2) + t = 2.4$. We can solve these two equations to find s and t . Subtracting equation 2 from equation 1: $(9/s + t) - (9/(s + 2) + t) = 4 - 2.4$. Simplify: $9/s - 9/(s + 2) = 1.6$. Factor out 9: $9[1/s - 1/(s + 2)] = 1.6$. Compute $1/s - 1/(s + 2) = (s + 2 - s)/[s(s + 2)] = 2/[s(s + 2)]$. So, $9 \times [2/(s(s + 2))] = 1.6$. Simplify: $18/[s(s + 2)] = 1.6$. Multiply both sides by $s(s + 2)$: $18 = 1.6s(s + 2)$. Divide both sides by 1.6: $18/1.6 = s(s + 2)$. Calculate $18/1.6 = 11.25$. So, $s(s + 2) = 11.25$. Which is: $s^2 + 2s - 11.25 = 0$. Quadratic equation: $s^2 + 2s - 11.25 = 0$. We can solve for s using quadratic formula: $s = [-2 \pm \sqrt{4 + 45}]/2$. Because discriminant is $2^2 - 4 \cdot 1 \cdot (-11.25) = 4 + 45 = 49$. So, $s = [-2 \pm 7]/2$. Since speed can't be negative, take the positive solution: $s = (-2 + 7)/2 = 5/2 = 2.5$ km/h. So, $s = 2.5$ km/h. Now, we can find t from equation 1: $9/s + t = 4$. $9/2.5 + t = 4$. $9/2.5$ is 3.6. So, $3.6 + t = 4$. Therefore, $t = 4 - 3.6 = 0.4$ hours. Convert 0.4 hours to minutes: $0.4 \times 60 = 24$ minutes. So, $t = 24$ minutes. Now, when she walks at $s + \frac{1}{2}$ km/h, which is $2.5 + 0.5 = 3$ km/h. We need to find the total time, which is walking time + t . Walking time is $9/3 = 3$ hours. Total time is 3 hours + 24 minutes. Convert 3 hours to minutes: 180 minutes. Total time: $180 + 24 = 204$ minutes. So, the walk takes 204 minutes, including the coffee shop time. **Final Answer** The walk takes 204 minutes.

ATTNPO-1.5B (ours):

Okay, let's try to solve this problem step by step. So, Aya walks 9 kilometers every morning. When she walks at a speed of s km/h, it takes her 4 hours, including t minutes in the coffee shop. When she increases her speed to $s + 2$ km/h, the total time becomes 2 hours and 24 minutes, still including t minutes in the coffee shop. We need to find the time when she walks at $s + \frac{1}{2}$ km/h. First, let's parse the information. When walking at speed s km/h: Distance = 9 km, Time taken walking + time in coffee shop = 4 hours. So, time walking is $\frac{9}{s}$ hours, and time in coffee shop is t minutes. But t is given in minutes, so we need to convert 4 hours to minutes to keep units consistent. 4 hours = 240 minutes. So, time walking + $t = 240$ minutes. Similarly, when walking at $s + 2$ km/h: Time taken walking + time in coffee shop = 2 hours 24 minutes = $2 \times 60 + 24 = 144$ minutes. So, time walking is $\frac{9}{s+2}$ hours, which is $\frac{9}{s+2} \times 60$ minutes, and time in coffee shop is t minutes. So, let's set up the equations. First case: $\frac{9}{s} \times 60 + t = 240$. Second case: $\frac{9}{s+2} \times 60 + t = 144$. We can subtract the first equation from the second to eliminate t . So, $[\frac{9}{s+2} \times 60 + t] - [\frac{9}{s} \times 60 + t] = 144 - 240$. Simplify: $\frac{9 \times 60}{s+2} - \frac{9 \times 60}{s} = -96$. Factor out 9×60 : $9 \times 60 \left(\frac{1}{s+2} - \frac{1}{s} \right) = -96$. Compute $\frac{1}{s+2} - \frac{1}{s} = \frac{s - (s+2)}{s(s+2)} = \frac{-2}{s(s+2)}$. So, $9 \times 60 \times \frac{-2}{s(s+2)} = -96$. Compute $9 \times 60 = 540$. So, $\frac{540 \times (-2)}{s(s+2)} = -96$. Multiply both sides by -1 : $\frac{540 \times 2}{s(s+2)} = 96$. So, $\frac{1080}{s(s+2)} = 96$. Thus, $s(s + 2) = \frac{1080}{96} = 11.25$. So, $s^2 + 2s - 11.25 = 0$. Solving quadratic equation: $s = \frac{-2 \pm \sqrt{4 + 45}}{2} = \frac{-2 \pm 7}{2}$. Since speed can't be negative, $s = \frac{5}{2} = 2.5$ km/h. So, $s = 2.5$ km/h. Now, find t . From first case: $\frac{9}{2.5} \times 60 + t = 240$. Compute $\frac{9}{2.5} = 3.6$. $3.6 \times 60 = 216$. So, $216 + t = 240 \Rightarrow t = 24$ minutes. Now, when walking at $s + \frac{1}{2} = 2.5 + 0.5 = 3$ km/h. Time walking: $\frac{9}{3} = 3$ hours = 180 minutes. Time in coffee shop: $t = 24$ minutes. Total time: $180 + 24 = 204$ minutes. **Final Answer** 204

Figure 13: This is a problem from AIME 2024, where TLMRE introduces unnecessary variables and follows an incorrect modeling idea before correcting itself, while ATTNPO avoids such detours and goes straight to the correct formulation, resulting in more concise reasoning.

Question (MMLU): A group of hikers climbed from Salt Flats (elevation -55 feet) to Talon Bluff (elevation 620 feet). What is the difference in elevation between Talon Bluff and Salt Flats? A. 565 feet ; B. 575 feet ; C. 665 feet ; D. 675 feet. Output the correct option (one of ABCD) in , such as .

TLMRE: Okay, so there's this problem about hikers climbing from Salt Flats to Talon Bluff, and I need to find the difference in elevation between these two places. Let me see... Salt Flats are at -55 feet, and Talon Bluff is at 620 feet. Hmm, the question is asking for the difference in elevation, so I think that means how much higher Talon Bluff is compared to Salt Flats. Wait, elevation can be positive or negative, so the difference would be the absolute value of the higher elevation minus the lower one. But let me clarify: the elevation of Salt Flats is -55 feet, which is below sea level, and Talon Bluff is at $+620$ feet, which is much higher. So the difference should be the total vertical distance between them. To find the difference, I can subtract the lower elevation from the higher one. So, 620 feet minus (-55) feet. Subtracting a negative is like adding, so that would be $620 + 55$. Let me calculate that: $620 + 55$ is 675 . So the difference in elevation is 675 feet. **Let me double-check: starting at -55 , you go up to 620 . The total change is $620 - (-55) = 620 + 55 = 675$. Yep, that seems right. Looking at the options, D is 675 feet. So the correct answer should be D. Final Answer .**

ATTNPO-1.5B(ours):

Okay, so I have this problem about hikers climbing from Salt Flats to Talon Bluff. The elevation of Salt Flats is -55 feet, and Talon Bluff is 620 feet. I need to find the difference in elevation between the two places. Hmm, let's see. Elevation difference is usually the absolute value of the difference between the two points. So, I should subtract the lower elevation from the higher one. But one is negative, so I need to handle that carefully. Elevation of Salt Flats is -55 feet, which is below sea level. Talon Bluff is 620 feet, which is much higher. So, the difference would be $620 - (-55)$. Subtracting a negative is like adding, so $620 + 55$ equals 675 feet. So the difference in elevation is 675 feet. Looking at the options, D is 675 feet. Final Answer .

Figure 14: This is a simple calculation problem from MMLU, and compared to ATTNPO, TLMRE exhibits redundant self-reflection.

Question (MMLU): Which of the jovian planets have rings? A. Neptune B. Uranus C. Saturn D. all of the above. Output the correct option (one of ABCD) in , such as .

TLMRE: Okay, so I need to figure out which jovian planets have rings. The options are Neptune, Uranus, Saturn, and all of the above. Let me start by recalling what I know about each of these planets. First, the jovian planets are the gas giants, right? So they're larger than the gas giants. The main ones are Jupiter, Saturn, Uranus, and Neptune. Now, I remember that gas giants do have rings, but I'm not entirely sure about each one. Let me think about each planet: Neptune: I believe Neptune has rings. It's a large planet, and like many gas giants, it has a ring system. **I think it's called the Great Ring System because it's a big, connected ring system around it.** Uranus: Uranus also has rings. It's another gas giant, and I recall that Uranus has a ring system that's quite prominent. There's a ring system that's pretty extensive around it. Saturn: Saturn is the largest of the gas giants. It has a ring system, but it's a bit more diffuse. **I think the rings are called Cassini's rings, which are not as tightly packed as Neptune's and Uranus's.** So all three: Neptune, Uranus, and Saturn have rings. Therefore, the correct answer should be D, all of the above.

ATTNPO-1.5B(ours):

Okay, so I have this question about jovian planets and their rings. The options are Neptune, Uranus, Saturn, and all of the above. Hmm, I remember that jovian planets are gas giants, so they're the gas giants themselves. Let me think about each option. Neptune: I recall that Neptune has a ring system. It's a gas giant with a lot of ice and rock around it. So Neptune has rings. Uranus: Similarly, Uranus is another gas giant. I think it also has a ring system. So Uranus also has rings. Saturn: Saturn is a gas giant too. It has rings as well. So Saturn has rings. Therefore, all three—Neptune, Uranus, Saturn—have rings. So the correct answer should be D, all of the above.

Figure 15: This is a simple knowledge-based question from MMLU. The response from TLMRE introduces information that is not directly necessary for solving the problem, whereas ATTNPO does not.