

# THINKTUNING: Instilling Cognitive Reflections without Distillation

Aswin RRV   Jacob Dineen   Divij Handa   Md Nayem Uddin  
Mihir Parmar   Chitta Baral   Ben Zhou

Arizona State University

{aravik13, chitta, xzhou202}@asu.edu

## Abstract

Recent advances in test-time scaling have led to the emergence of thinking LLMs that exhibit self-reflective behaviors and multi-step reasoning. While RL drives this self-improvement paradigm, a recent study (Gandhi et al., 2025) shows that RL alone does not truly instill these new reasoning abilities - it merely draws out behaviors already present in the base models. This raises a question: *How can we train models that don't exhibit such thinking behavior to develop it in the first place?* To this end, we propose THINKTUNING, a GRPO-based interactive training approach where we augment the rollouts of a student model with the guidance from a teacher model. A simple idea from classroom practice inspires our method: a teacher poses a problem, lets the student try an answer, then gives corrective feedback—enough to point the mind in the right direction and then show the solution. Each piece of feedback reshapes the student's thoughts, leading them to arrive at the correct solution. Similarly, we find that this type of implicit supervision through feedback from a teacher model of the same size improves the reasoning capabilities of the student model. In particular, on average, our method shows a 3.85% improvement over zero-shot baselines across benchmarks, and on MATH-500, AIME and GPQA-Diamond it shows 2.08%, 2.23% and 3.99% improvements over the vanilla-GRPO baseline<sup>1</sup>.

## 1 Introduction

Recent progress in AI research has been driven by advances in scaling the models' parameter count (Kaplan et al., 2020). More recently, scaling along the inference-time axis has produced significant performance gains in various complex reasoning tasks (Snell et al., 2025). Thinking models such as OpenAI-o-series (Jaech et al., 2024), DeepSeek-R1 (Guo et al., 2025) and Gemini-Thinking (Team

<sup>1</sup>Source code is available at <https://github.com/3rdAT/ThinkTuning>

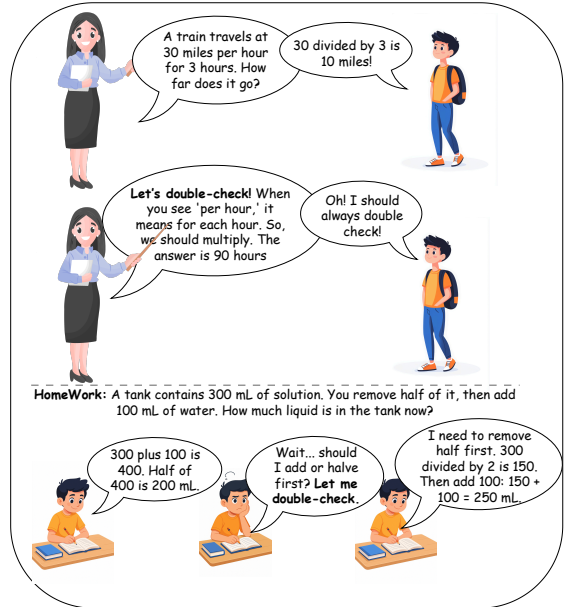


Figure 1: Illustration of THINKTUNING motivation. Top: teacher poses a math problem, student answers incorrectly, and the teacher offers a short corrective feedback. Bottom: For a new problem, the student recalls the feedback ("double-check") and ends up producing the correct answer.

et al., 2023) are a testament to this, capable of producing long reasoning chains, with sophisticated behaviors like self-reflection, self-correction, and multi-step reasoning. These significant performance gains are attributed to the success of Reinforcement Learning (RL) through simple rule-based rewards. However, online on-policy RL settings face a constraint: sophisticated reasoning behaviors will not emerge unless they are explicitly sampled during training. For example, models like Qwen (Yang et al., 2025) often come with strong priors, allowing them to naturally generate sophisticated reasoning behaviors, which RL then amplifies. In contrast, when models lack strong priors, on-policy RL struggles to elicit them. Indeed, a recent study shows that RL applied on the

Llama 3.2-family (Grattafiori et al., 2024) of models struggles to elicit sophisticated reasoning behaviors (Gandhi et al., 2025).

In academic settings, cognitive modeling provides a structured approach for shaping both overt (external) and covert (internal-cognitive) behaviors of students through guided interventions by a teacher, typically using verbal mediation (Camp and Bash, 1978). As illustrated in Fig. 1, imagine a teacher asking, “A train travels at 30 miles per hour for 3 hours. How far does it go?” A hasty student might reply, “30 divided by 3 is 10 miles!” A good teacher not only explains why the answer is incorrect but also imparts a generalizable skill. In this case, the teacher could encourage the student to double-check what “per hour” means and to think carefully about whether they should multiply or divide in similar problems. Particularly, in STEM education literature (Chouvalova et al., 2024), it has been established that the use of corrective feedback with errorful learning is more beneficial in student learning<sup>2</sup>. Interestingly, recent thinking models often exhibit such behavior of re-checking and self-refining, which makes them better at various reasoning tasks. Presumably, these thinking behaviors emerge in those models solely through RL, as suitable priors are present to help in exhibiting such behavior (Gandhi et al., 2025). However, this brings up an important question: *How can we enable models to acquire these types of thinking skills in the absence of suitable priors? And is RL alone sufficient for this task?*

Drawing inspiration from the example discussed above, we propose THINKTUNING, a training approach where an active student model learns to think by interacting with a teacher model. Rather than assuming thinking behaviors will emerge during RL, we engineer the training process to induce them. This aligns with how cognitive modeling in educational settings elicits complex reasoning strategies, such as self-reflection, self-correction, and problem-solving among students.

THINKTUNING consists of two stages. First, we start by creating a set of few-shot exemplars, each demonstrating an opinion on a student’s response, a reason for that opinion, and a phrase that typically showcases specific cognitive behaviors by solving a problem. Our exemplars capture the most common

human self-reflective behaviors: Self-Conflict, Self-Agreement, Self-Critique, and Self-Consultancy. While many other cognitive behaviors exist, we focus on these four because they are well defined (Hermans, 2023; Hermans and Gieser, 2011). Second, we train the student model in an online RL setting with Group Relative Policy Optimization (GRPO) (Shao et al., 2024). At each iteration, the student model generates  $n$  rollouts, from which a subset of  $\gamma$  rollouts is randomly selected. These selected rollouts are passed to the few-shot teacher model to obtain feedback and phrases showcasing the cognitive skill. The feedback is then appended to the corresponding  $\gamma$  rollouts. The resulting  $\gamma_{aug}$  rollouts, together with the remaining  $n - \gamma_{aug}$  un-augmented rollouts, are used for computing the advantage estimates for the GRPO algorithm.

However, because the teacher model’s guidance is entirely off-policy, it violates the assumptions required for importance sampling in GRPO. To address this, we introduce Advantage-Aware Shaping (AAS), which adjusts the updates for tokens generated with teacher guidance by taking into account both the advantage and the student model’s current confidence in producing the token. This helps prevent unstable updates during training and keeps the model from becoming degenerate.

Our experiments show that a model trained with THINKTUNING improve performance across diverse reasoning benchmarks like GSM8k (+3.14%), MATH-500 (+9.4%), AIME (+4.94%), CSQA (+3.04%), ARC-Challenge (+4.31%), GPQA-Diamond (+3.08%) and MMLU-Pro (+2.8%) compared to zero-shot baselines. Our training approach improves over the GRPO baseline by 2.08%, 2.23% and 3.99% on MATH-500, AIME and GPQA-Diamond respectively. Our analysis experiment showcases that THINKTUNING can steer the exploration during RL training and instill unknown behaviors in the policy model.

## 2 Related Works

**Inference-Time Scaling** Scaling inference-time compute has been a promising approach to improve LLMs’ performance. Chain-of-thought (CoT) encourages models to generate step-by-step reasoning, significantly boosting performance on complex tasks (Wei et al., 2022; Kojima et al., 2022). Self-consistency generates multiple reasoning paths and selects the most frequent answer, further improving accuracy (Wang et al., 2023). Iterative self-

<sup>2</sup>A Quote from (Chouvalova et al., 2024): “Compared to passively reading materials, errorful learning paired with corrective feedback is more beneficial to student learning and retention (Mera et al., 2022; Overman et al., 2021).”

refinement, where models critique and correct their own outputs, yields additional gains without weight updates (Madaan et al., 2023). Methods such as Tree-of-Thoughts and MCTS<sub>r</sub> extend inference-time search by exploring branching reasoning trajectories (Yao et al., 2023). Recent works, test-time optimization (Snell et al., 2025) and PlanGEN (Parmar et al., 2025b), put emphasis on dynamically adjusting inference compute based on the complexity of the task. Recently, Xu et al. (2024) view the next-token prediction as a fundamental reasoning task, and proposes annotating pretraining texts by explaining why a particular next word should follow and how it connects to the preceding context. By continually pretraining on this augmented data, they demonstrate that the reasoning abilities of LLMs improve. In contrast to all these approaches, our work focuses on training models to increase their inference-compute during test time by instilling cognitive reflections in their responses.

**Online and Offline RL** Online RL involves a model interacting with an environment to obtain rewards and updating its parameters to maximize them. Proximal Policy Optimization (PPO) underpins most RLHF pipelines, aligning LLMs to human preferences (Schulman et al., 2017; Ouyang et al., 2022) in an online way. In contrast, Offline RL involves making use of pre-collected data, such as preference-labeled datasets. Directive Preference Optimization (DPO) (Rafailov et al., 2023) reformulates preference alignment as a supervised objective, matching or outperforming PPO in stability and quality. Variants of DPO use three preferences instead of two, showing better performance on reasoning tasks (Saeidi et al., 2024). A recent variant of PPO, Group Relative Policy Optimization (GRPO) (Shao et al., 2024) discards the critic network from PPO and computes the advantage estimates by comparing each trajectory’s reward to the mean reward of a group of sampled trajectories, thus improving efficiency and scalability of RL training. This has been effective in improving the reasoning and planning capabilities of LLMs (Parmar et al., 2025a). Our work is different from these approaches as we try to obtain off-policy guidance during online RL training.

**Off-Policy Guidance during RL** Earlier works in RL like (Schmitt et al., 2018) showcase that kickstarted training improves the data efficiency of agents being trained. Kickstarting demonstrated up to 10x faster training and convergence of the

agents. Recent work done by Yan et al. (2025) closely aligns with our work. The authors include samples from a larger model, such as Deepseek-R1, alongside the on-policy rollouts during GRPO. They propose Policy Shaping, which corrects the importance-sampling ratios during training. However, our work differs from theirs by dynamically calculating the shaping coefficient and augmenting on-policy rollouts with off-policy tokens.

## 3 Methods

### 3.1 Background

**GRPO** The recent success of DeepSeek-R1 (Guo et al., 2025) has established GRPO as the preferred algorithm for online reinforcement learning due to its efficiency and ease of implementation. GRPO, a PPO (Schulman et al., 2017) variant, estimates the advantage by aggregating the reward scores of a group of  $n$  sampled responses to a given query  $q$ , thus eliminating the need for a separate value network and generalized advantage estimation (GAE) (Schulman et al., 2015). Formally, let  $\mathcal{M}_\theta$  and  $\mathcal{M}_{\theta_{old}}$  be the current and old policy models, respectively. Let  $q$  and  $o_i$  be the query and  $i^{th}$  response sampled from the dataset and the old policy, respectively. Let  $r(\cdot)$  be the reward function, which measures the correctness of a given response. Then, the GRPO objective is defined as follows:

$$\begin{aligned} \mathcal{J}_{GRPO}(\theta) = \mathbb{E} \Big[ & q \sim \mathcal{D}, \{o_i\}_{i=1}^n \sim \mathcal{M}_{\theta_{old}}(O | q) \Big] \\ & \left\{ \frac{1}{n} \sum_{i=1}^n \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min \left[ \frac{\mathcal{M}_\theta(o_{i,t}|q, o_{i,<t})}{\mathcal{M}_{\theta_{old}}(o_{i,t}|q, o_{i,<t})} \hat{A}_{i,t}, \right. \right. \\ & \left. \text{clip} \left( \frac{\mathcal{M}_\theta(o_{i,t}|q, o_{i,<t})}{\mathcal{M}_{\theta_{old}}(o_{i,t}|q, o_{i,<t})}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right] \\ & \left. - \beta D_{KL}[\mathcal{M}_\theta \parallel \mathcal{M}_{ref}] \right\} \end{aligned}$$

Here, the advantage is calculated as the normalized reward, i.e.,  $\hat{A}_{i,t} = \tilde{r}(o_i) = \frac{r(o_i) - \text{mean}(r)}{\text{std}(r)}$ . This eliminates the need for complicated advantage estimation that happens in PPO. In the above expression,  $\frac{\mathcal{M}_\theta(o_{i,t}|q, o_{i,<t})}{\mathcal{M}_{\theta_{old}}(o_{i,t}|q, o_{i,<t})}$ , is the importance sampling weight which corrects for the mismatch between the current policy  $\mathcal{M}_\theta$  and the old policy  $\mathcal{M}_{\theta_{old}}$  that generated the sample responses. This importance sampling weight ( $w$ ) ensures that updates are properly reweighted so that learning remains unbiased even when the policy changes over the course of training.

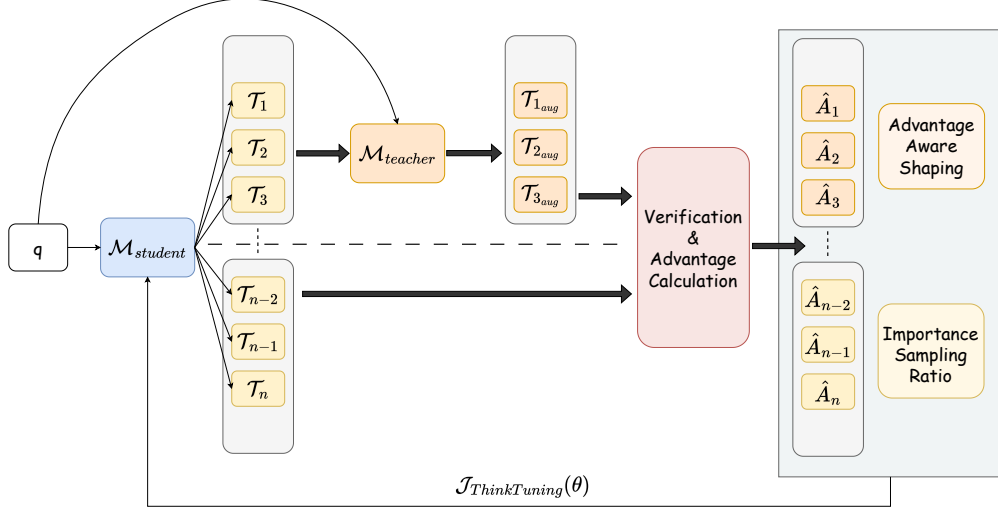


Figure 2: **ThinkTuning**: The student model  $\mathcal{M}_{\text{student}}$  generates  $n$  rollouts  $T_1, \dots, T_n$  for question  $q$ . A selected subset (e.g.  $T_1, T_2, T_3$ ) is passed (along with  $q$ ) to the teacher model  $\mathcal{M}_{\text{teacher}}$ , producing augmented rollouts  $\mathcal{T}_{\text{aug}}$ . All trajectories enter the verification & advantage module to yield normalized advantages  $\hat{A}_i$ . Augmented tokens are weighted via Advantage Aware Shaping; remaining tokens use the standard importance sampling ratio. These per-token weights are used in  $\mathcal{J}_{\text{ThinkTuning}}(\theta)$  for updating the student.

## 3.2 THINKTUNING

### 3.2.1 Student Responses (*student responds*)

In the first stage of THINKTUNING, we sample  $n$  responses from the student policy  $\mathcal{M}_{\text{student}}$  for each query  $q$  in a training batch drawn from the dataset  $\mathcal{D}$ . We sample responses at a temperature of 1.0 to observe diversity. These initial  $n$  responses represent the student model’s unaided attempts at solving a given problem, typically exhibiting a mix of correct, partially correct, and incorrect reasoning.

### 3.2.2 Teacher Guidance (*teacher helps*)

In the second stage, we obtain guidance from the teacher model  $\mathcal{M}_{\text{teacher}}$ . Given the student model’s response, the teacher model provides its guidance by first stating its opinion. Then, it provides its justification for its opinion, grounded in its own reasoning process, and finally offers a guiding phrase on how to approach and solve the problem effectively. Throughout this process, the teacher model explicitly demonstrates cognitive behaviors, serving as an exemplar of reflective problem-solving strategies for the student to learn from. In particular, we focus on four self-reflective cognitive behaviors, well-defined in (Hermans, 2023; Hermans and Gieser, 2011): (1) **Self-Conflict**: challenging one’s own response by presenting alternative perspectives; (2) **Self-Critique**: identifying weaknesses in their response and suggesting improvements; (3) **Self-Agreement**: affirming and

justifying the strengths in their response; and (4) **Self-Consultancy**: drawing on an alternative internal perspective or source of expertise to offer new advice or insights that could further improve one’s own response. We provide four few-shot exemplars—two illustrating incorrect student responses and two showcasing correct ones—each demonstrating one of the mentioned behaviors. Importantly, all exemplars are expressed in the first-person perspective, framing the guidance as inner dialogue or self-reflection, making it natural for the student model to imitate during training.

After obtaining the rollouts for a given query from the student model, we randomly pass a fraction  $\gamma$  of the rollouts to the teacher model for guidance. For each selected rollout  $o_i$ , we give the corresponding question  $q$  to the teacher model  $\mathcal{M}_{\text{teacher}}$ . With the help of our few-shot exemplars, we obtain the guidance from the teacher model in a structured way, as shown in Appendix A.1.

### 3.2.3 Student Training (*student improves*)

In this stage, the feedback generated by the teacher model ( $\mathcal{M}_{\text{teacher}}$ ) is augmented to the selected fraction  $\gamma$  of the corresponding student rollouts. This produces a set of  $\gamma_{\text{aug}}$  augmented trajectories. These are combined with the remaining  $n - \gamma_{\text{aug}}$  un-augmented student rollouts to compute token-level advantage estimates used in the GRPO update. We formally call this process



$\text{Guide}(\mathcal{M}_{\text{teacher}}, \mathcal{M}_{\text{student}_{\theta_{\text{old}}}}, q, \gamma)$ . Then, we compute the group-normalized advantage for each token in a trajectory  $\mathcal{T}_i \in \{\mathcal{T}_{\text{unaug}} \cup \mathcal{T}_{\text{aug}}\}$  as:

$$\hat{A}_{i,t} = \tilde{r}(\mathcal{T}_i) = \frac{R(\mathcal{T}_i) - \text{mean}(\mathcal{R}(\mathcal{T}_{\text{unaug}} \cup \mathcal{T}_{\text{aug}}))}{\text{std}(\mathcal{R}(\mathcal{T}_{\text{unaug}} \cup \mathcal{T}_{\text{aug}}))}$$

Here,  $\mathcal{T}_{\text{unaug}}$  denotes the set of unaugmented trajectories, and  $\mathcal{T}_{\text{aug}}$  denotes the teacher-augmented ones. When teacher guidance successfully reasons toward the correct answer, the augmented trajectory typically receives a higher reward, resulting in a higher relative advantage. In contrast, if the guidance is not helpful, the unaugmented trajectories dominate the normalization, which automatically reduces the effect of poor teacher interventions.

---

**Algorithm 1** THINKTUNING

---

```

1: Input: Initial Student model  $\mathcal{M}_{\text{student}_{\theta_{\text{init}}}}$ , Teacher
   model  $\mathcal{M}_{\text{teacher}}$ , guidance fraction  $\gamma$ , hyperparameter
   set  $(\epsilon, \beta, c_1, c_2, k)$ 
2:
3:  $\mathcal{M}_{\text{student}_{\theta}} \leftarrow \mathcal{M}_{\text{student}_{\theta_{\text{init}}}}$ 
4:
5: for training step=1 to  $I$  do
6:    $\mathcal{M}_{\text{student}_{\text{old}}} \leftarrow \mathcal{M}_{\text{student}_{\theta}}$ 
7:   Sample batch  $\mathcal{D}_b \subset \mathcal{D}$ 
8:
9:   // Student acts & Teacher helps
10:  for all  $q \in \mathcal{D}_b$  do
11:    if training step  $\leq k$  then
12:       $\{o\}_{i=1}^n \sim \text{Guide}(q, \mathcal{M}_{\text{student}_{\text{old}}}, \mathcal{M}_{\text{teacher}}, \gamma)$ 
13:    else
14:       $\{o\}_{i=1}^n \sim \mathcal{M}_{\text{student}_{\text{old}}}(O | q)$ 
15:    end if
16:  end for
17:
18:  // Reward calculation and Advantage estimation
19:  Compute the rewards  $r_i = r(o_i)$  for each response
   Compute group-normalized advantage  $\hat{A}_{i,t}$  for all to-
   kens
20:
21:  for mini-batch step = 1 to  $\mu$  do
22:    if training step  $\leq k$  and  $o_i \in \mathcal{T}_{\text{aug}}$  then
23:      Calculate  $w_{\text{aas}}(o_i)$ 
24:    else
25:      Calculate  $w(o_i)$ 
26:    end if
27:     $\mathcal{M}_{\text{student}_{\theta}} \leftarrow \arg\max_{\theta} \mathcal{J}_{\text{THINKTUNING}}(\theta)$ 
28:  end for
29: end for
30: Output: Final think-tuned model  $\mathcal{M}_{\text{student}_{\theta}}$ 

```

---

**Off-policy guidance tokens** A core challenge arises from the fully off-policy nature of the tokens from teacher guidance. Although importance sampling can, in principle, correct for the distributional mismatch, accurate correction would require access to  $\mathcal{M}_{\text{teacher}}(\text{guidance} | q, o_{\text{student}})$ . In practice, however, this does not reflect the true probability with which the guidance was sampled from the teacher model,

due to differences in the prompting setup. To address this, we propose Advantage-Aware Shaping (AAS) for the augmented tokens in the trajectories  $\mathcal{T}_{\text{aug}}$  instead of using the importance sampling weights. AAS uses the student model’s own confidence in the tokens of the augmented trajectory, modulated by its relative advantage, to determine the weight assigned to each teacher-injected token’s gradient during training. Formally, for each augmented off-policy token  $o_t^{\text{aug}}$  in the trajectory  $\mathcal{T}_{\text{aug}}$ , we define the Advantage Aware Shaping (AAS) weight as:

$$w_{\text{aas}}(o_t^{\text{aug}}, \hat{A}_t) = \frac{\mathcal{M}_{\text{student}}(o_t^{\text{aug}} | q, o_{<t})}{sg(\mathcal{M}_{\text{student}}(o_t^{\text{aug}} | q, o_{<t})) + c(\hat{A}_t)}$$

where  $sg$  denotes the stop-gradient operator and  $\mathcal{M}_{\text{student}}(o_t^{\text{aug}} | q, o_{<t})$  denotes the probability assigned by the student model to the token  $o_t^{\text{aug}}$ , given the query  $q$  and the preceding tokens  $o_{<t}$ . This formulation is similar to the policy shaping proposed by Yan et al. (2025). However, in THINKTUNING we make use of  $c(\hat{A}_t)$ , a shaping coefficient determined by the advantage  $\hat{A}_t$  at that token. To be specific,  $c(\hat{A}_t)$  is computed as:

$$c(\hat{A}_t) = c_1 + (c_2 - c_1) \cdot \frac{A_{\text{max}} - \hat{A}_t}{A_{\text{max}} - A_{\text{min}}}$$

where  $c_1$  and  $c_2$  are hyperparameters, and  $A_{\text{min}}$ ,  $A_{\text{max}}$  are the minimum and maximum token advantages possible for a group of responses. This is a linear mapping function which provides a shaping coefficient close to  $c_1$  for positive advantages and a shaping coefficient close to  $c_2$  for negative advantages. These shaping coefficients serve as a knob that enables us to control the magnitude of gradient updates for the off-policy tokens. For a detailed analysis of its effect on  $w_{\text{aas}}$  and its subsequent impact on gradient updates, see Appendix A.3.

**Training Objective** We incorporate this shaping mechanism directly into our final training objective, which we refer to as  $\mathcal{J}_{\text{THINKTUNING}}(\theta)$ . For each of the on-policy tokens  $o_t \in \{\mathcal{T}_{\text{unaug}} \cup \mathcal{T}_{\text{aug}}\}$  in the batch, we compute the importance sampling weight  $w_t$  between the current and old student policies. For off-policy tokens in the teacher-augmented part of the trajectories, i.e.,  $o_t^{\text{aug}} \in \mathcal{T}_{\text{aug}}$ , we make use of the advantage-aware shaped weight  $w_{\text{aas}}$  as discussed above. In order to distinguish between the on-policy student tokens ( $o_t$ ) and the off-policy guidance tokens ( $o_t^{\text{aug}}$ ) during loss computation, we utilize a binary mask ( $m_t$ ). Formally, we define the THINKTUNING objective as follows:

$$\begin{aligned}
\mathcal{J}_{\text{THINKTUNING}}(\theta) = & \mathbb{E} \left[ q \sim \mathcal{D}, \{o_i\}_{i=1}^n \sim \text{Guide}(q, \mathcal{M}_{\theta_{\text{old}}}, \mathcal{M}_{\text{teacher}}, \gamma) \right] \\
& \left\{ \frac{1}{n} \sum_{i \in \mathcal{T}_{\text{unaug}}} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min \left[ w_{i,t} \hat{A}_{i,t}, \right. \right. \\
& \left. \left. \text{clip}(w_{i,t}, 1 - \epsilon, 1 + \epsilon) \hat{A}_{i,t} \right] \right. \\
& + \frac{1}{n} \sum_{i \in \mathcal{T}_{\text{aug}}} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left[ m_{i,t} \cdot w_{\text{aas}}(o_{i,t} \hat{A}_{i,t}) \cdot \hat{A}_{i,t} + \right. \\
& \left. (1 - m_{i,t}) \cdot (\min[w_{i,t} \cdot \hat{A}_{i,t}, \text{clip}(w_{i,t}, 1 - \epsilon, 1 + \epsilon) \cdot \hat{A}_{i,t}]) \right] \\
& \left. - \beta D_{KL}[\mathcal{M}_{\theta} \parallel \mathcal{M}_{\text{ref}}] \right\}
\end{aligned}$$

where  $w$  and  $w_{\text{aas}}$  are importance sampling and advantage-aware shaped weights, respectively. This formulation preserves the benefits of GRPO’s group-relative advantage estimation while addressing the off-policy nature of teacher-augmented rollouts through controlled shaping. As a result, the student model is encouraged to learn from helpful feedback of the teacher model. Once the student model has sufficiently learned from the teacher’s guidance, we stop providing further guidance after a predefined number of steps ( $k$ ), which is a hyperparameter.

## 4 Experiments

### 4.1 Setup

**Baselines** We first compare our method against zero-shot baselines and prompt-based self-improvement methods like Self-Verify (Kumar et al.) and Self-Correct (Huang et al., 2023). We also compare with the s1-budgeting (Muennighoff et al., 2025) method, where we set a token budget of 2048 and let the model generate until it reaches this budget by replacing the end-of-sequence token with “wait...”. For training-based methods, we compare against SFT, STaR (as implemented by Kumar et al.), and GRPO (Guo et al., 2025).

**Training Dataset** For THINKTUNING and other training-based methods, we make use of the GSM8k train set which has 7473 samples. We train only on this dataset to showcase that THINKTUNING could generalize to out-of-domain problems.

**Models** We use **Llama3.2-3B-Instruct** (Grattafiori et al., 2024) model as the base model to obtain our baselines and perform training with THINKTUNING. Recent work (Gandhi et al., 2025) shows that models like Qwen naturally exhibit these cognitive behaviors, whereas the Llama family of models lacks them. Hence, choosing a

model from the Llama family is a natural way to demonstrate the utility of our method. We also use the same 3B version for the teacher model.

**Benchmarks** We evaluate our method on several benchmarks across different reasoning categories: **GSM8K** (Cobbe et al., 2021), **MATH-500** (Hendrycks et al., 2021) and **AIME** (Veeraboina, 2023) for Mathematical Reasoning; **CSQA** (Talmor et al., 2018) and **StrategyQA** (Geva et al., 2021) for Commonsense Reasoning; and for Scientific Reasoning, we use **ARC-Challenge** (ARC-C) (Clark et al., 2018) and **GPQA Diamond Set** (GPQA-D) (Rein et al., 2024) (see Table 1). To ensure consistent and proper evaluation, after the model finishes generation, we append the phrase “So, the final answer is \boxed{” , which prompts the model to explicitly output the final answer in a boxed format, simplifying answer parsing and enabling exact match (EM) accuracy calculation using Math-Verify with ease.

**Training & Inference** We implement THINKTUNING using the ver1 (Sheng et al., 2024) framework. All experiments are conducted on 4 NVIDIA H100 GPUs. For detailed hyperparameter settings, please refer to the Appendix A.2. To speed up rollout generation and evaluation, we utilize vLLM (Kwon et al., 2023) due its efficiency.

### 4.2 Results

**Comparison with prompting-based methods** From Table 1, we can see that Self-Verify and Self-Correct methods underperform compared to the Zero-Shot-CoT baseline. They achieve only 52.08% and 51.45% on GSM8k and 34.98% and 32.46% on Math-500, respectively, whereas Zero-Shot-CoT attains 71.08% and 38.14% on these benchmarks. We see similar trends on other benchmarks like CSQA, ARC-C, GPQA-D and MMLU-Pro. The s1-budgeting method, which simply scales inference-time compute, yields only marginal improvements on GPQA-D yet remains far below the baseline on other reasoning tasks. Our evaluation shows that this method fails to produce meaningful gains, and in several cases leads to degraded performance. For instance, on MATH-500, s1-budgeting yields only 25.72%, underperforming even the Zero-Shot-CoT baseline, and on CSQA, it performs on par with Self-Verify but remains 16.2 points behind THINKTUNING (54.21% vs. 70.43%). In contrast, our THINKTUNING consistently outperforms Zero-Shot-CoT

Methods	Mathematical Reasoning			CommonSense Reasoning	Scientific Reasoning		Other Reasoning	
	GSM8K	MATH-500	AIME	CSQA	ARC-C	GPQA-D	STRATEGYQA	MMLU-PRO
Zero-Shot-CoT	71.08 $\pm$ 0.20	38.14 $\pm$ 0.75	9.32 $\pm$ 0.36	67.39 $\pm$ 0.26	75.49 $\pm$ 0.20	25.10 $\pm$ 0.85	66.40 $\pm$ 0.43	34.41 $\pm$ 0.11
Self-Verify	52.08 $\pm$ 1.73	34.98 $\pm$ 0.54	8.19 $\pm$ 0.29	54.41 $\pm$ 0.73	61.56 $\pm$ 0.47	23.94 $\pm$ 0.68	52.10 $\pm$ 0.39	28.10 $\pm$ 0.14
Self-Correct	51.45 $\pm$ 0.30	32.46 $\pm$ 0.47	7.81 $\pm$ 0.18	45.90 $\pm$ 0.69	52.88 $\pm$ 0.58	24.60 $\pm$ 0.71	52.39 $\pm$ 0.78	25.50 $\pm$ 0.12
s1-budgeting	51.30 $\pm$ 0.42	25.72 $\pm$ 0.54	9.01 $\pm$ 0.31	54.21 $\pm$ 0.44	59.51 $\pm$ 0.27	26.57 $\pm$ 0.99	57.88 $\pm$ 0.80	28.59 $\pm$ 0.10
SFT	62.27 $\pm$ 0.61	29.00 $\pm$ 0.49	6.07 $\pm$ 0.43	65.91 $\pm$ 0.24	70.90 $\pm$ 0.71	24.49 $\pm$ 0.82	64.12 $\pm$ 0.65	36.07 $\pm$ 0.07
STaR	73.54 $\pm$ 0.22	40.78 $\pm$ 0.35	8.91 $\pm$ 0.29	67.91 $\pm$ 0.30	77.24 $\pm$ 0.21	21.46 $\pm$ 0.86	66.84 $\pm$ 0.41	34.69 $\pm$ 0.12
GRPO	<b>78.89<math>\pm</math>0.84</b>	45.46 $\pm$ 1.55	12.03 $\pm$ 0.33	69.86 $\pm$ 0.52	79.13 $\pm$ 0.21	24.19 $\pm$ 0.75	<b>70.68<math>\pm</math>0.35</b>	36.07 $\pm$ 0.07
THINKTUNING	74.22 $\pm$ 0.13	<b>47.54<math>\pm</math>0.46</b>	<b>14.26<math>\pm</math>0.38</b>	<b>70.43<math>\pm</math>0.19</b>	<b>79.80<math>\pm</math>0.24</b>	<b>28.18<math>\pm</math>0.63</b>	66.52 $\pm$ 0.41	<b>37.21<math>\pm</math>0.11</b>

Table 1: **Main Results.** We evaluate eight methods on *seven* benchmarks that we group into a four-way taxonomy: (i) *Mathematical reasoning* (GSM8K, MATH-500); (ii) *Commonsense reasoning* (CSQA); (iii) *Scientific reasoning* (ARC-CHALLENGE, GPQA-DIAMOND); and (iv) *Other multi-disciplinary reasoning* (STRATEGYQA, MMLU-PRO). We report accuracy (%) as the mean  $\pm$  standard error over ten random seeds. For each dataset the highest score is **boldfaced** and the second-highest is underlined. All experiments were run with a maximum context length of 4096 tokens and a decoding temperature of 0.7.

and all prompt-based methods. It achieves 74.22% on GSM8k (+3.14 points), 47.54% on Math-500 (+9.40 points), and similar gains on CSQA, ARC-C, GPQA-D, StrategyQA, and MMLU-Pro.

**Comparison with training-based methods** Our experiments show that fine-tuning (SFT) on the GSM8k training split degrades performance across every benchmark. Interestingly, we also observe that SFT leads to a performance drop of around 8% even on the GSM8k test set. We hypothesize that this is due to a distributional mismatch between the Llama 3.2 family’s pretrained reasoning priors and the highly structured chain-of-thought formats found in the GSM8k training annotations. In contrast, the STaR method, which uses the self-generated reasoning chains into the fine-tuning process, achieves 73.54 % on GSM8k (vs. 62.27 % for SFT) and 40.78 % on Math-500 (vs. 29.00 %). It also improves on CSQA (67.91 % vs. 65.91 %) and ARC-C (77.24 % vs. 70.90 %), but its gains are uneven: STaR scores only 21.46 % on GPQA-D and records 66.84 % on StrategyQA and 34.69% on MMLU-Pro. By comparison, THINKTUNING consistently outperforms STaR across all benchmarks—74.22 % on GSM8k (+0.68 points), 47.54 % on Math-500 (+6.76 points), 70.43 % on CSQA (+2.52 points), 79.80 % on ARC-C (+2.56 points), and 28.18 % on GPQA-D (+6.72 points).

**Comparison with GRPO** GRPO serves as our strongest online RL baseline, and achieves 78.89 % on GSM8k, 45.46 % on Math-500, 69.86 % on CSQA, 79.13 % on ARC-C, and 24.19 % on GPQA-D. On broader reasoning tasks, GRPO attains 70.68 % on StrategyQA and 36.07 % on MMLU-Pro. In comparison, THINKTUNING underperforms GRPO on GSM8k (74.22% vs. 78.89

%) and StrategyQA (66.52 % vs. 70.68 %) but outperforms on rest: Math-500 (47.54 % vs. 45.46 %), CSQA (70.43 % vs. 69.86 %), ARC-C (79.80 % vs. 79.13 %), and GPQA-D (28.18 % vs. 24.19 %). Also, THINKTUNING exceeds GRPO on MMLU-Pro (37.21 % vs. 36.07 %), demonstrating stronger scientific and factual reasoning.

## 5 Analysis

**Does THINKTUNING scale inference time?** To investigate this, we analyze the number of tokens generated during our evaluation. Specifically, we compare the output length of responses from models trained with GRPO and THINKTUNING across six benchmarks, excluding AIME and MMLU-Pro. For each benchmark, we compute the average number of tokens generated per question and report the results in Figure 3. We observe that both GRPO and THINKTUNING models end up spending more compute on complex benchmarks that require multi-step reasoning. For example, in benchmarks like MATH-500 and GPQA-D, they produce responses with more than 300 tokens. However, on the GPQA-D benchmark THINKTUNING model ends up spending around 5.2% more tokens than the GRPO-trained model, which translates into an improvement in relative performance. Interestingly, the GRPO model spends 3.6% more tokens than THINKTUNING model, but the latter ends up performing better in MATH-500. On other benchmarks as well, THINKTUNING model spends around 3.4-20.8% more tokens than the GRPO model. From these analyses, it is evident that THINKTUNING increases inference-time compute by instilling cognitive reflection, which results in performance improvements in certain benchmarks.

GSM8k Error Categories			StrategyQA Error Categories		
Error Type	Description	Freq	Error Type	Description	Freq
Computation Errors	Mistakes in arithmetic, algebra, or basic number calculations	45%	Knowledge-Retrieval Errors	Failure to fetch or recognize the correct factual premises.	65%
Interpretation Errors	Misreading the question or using the wrong quantities/units	85%	Interpretation Errors	Misreading the question and interpreting it wrongly.	15%
Logical Reasoning Errors	Faulty step-by-step logic, including contradictions or invalid inferences	70%	Logical-Inference Errors	Faulty reasoning, including contradictions or invalid inferences.	20%
Recall Errors	Forgetting earlier facts or intermediate results already computed	15%	Answer Label Errors	Sound reasoning with correct premises, but incorrect Boolean label.	5%
Redundancy Errors	Unnecessary steps or checks that increase solution length and introduce mistakes	40%			

Table 2: **Error Category Descriptions and Frequencies.** Side-by-side comparison of error types, their descriptions, and frequency of occurrence observed in GSM8K and STRATEGYQA analyses (based on 20 sampled instances where GRPO model was correct and THINKTUNING model was incorrect).

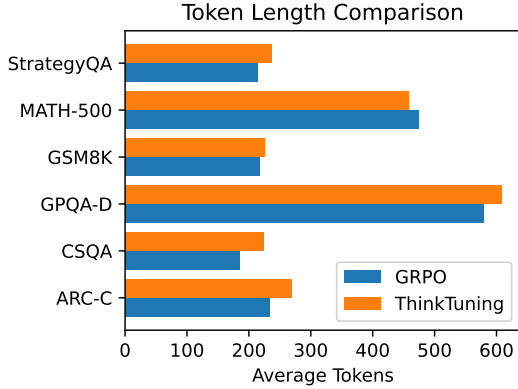


Figure 3: Average number of tokens generated per question by models trained with GRPO and THINKTUNING across six reasoning benchmarks (StrategyQA, MATH-500, GSM8K, GPQA-D, CSQA, and ARC-C).

### Error Analysis on GSM8K and StrategyQA

As observed in Table 1, GRPO has better performance than THINKTUNING on the GSM8k and StrategyQA benchmarks. We further investigate this performance gap by conducting a manual error analysis on 20 instances where the GRPO model was correct and the THINKTUNING model was incorrect. Our analysis reveals that the THINKTUNING model exhibits different types of errors. These error types along with their distribution across the analyzed instances are shown in Table 2. On GSM8k instances, we observe that errors arising from misinterpretation of the question occur in 85% of the analyzed instances. Furthermore, such misinterpretation often leads to logical inconsistencies, which occur in 70% of the instances. Upon closer examination, we observed that Self-Conflict and Self-Critique strategies lead the model to second-guess correct initial steps, eventually ending up misinterpreting the problem. Similarly, in StrategyQA instances, we observe that Knowledge Retrieval errors occur in 65% of the instances. Here, most of the time, the model agrees with incorrect knowl-

edge recalled and ends up providing the wrong answer. However, these strategies prove to be beneficial in much harder math problems like Math-500, AIME, GPQA-D, MMLU-Pro, but hinder performance in simple benchmarks like GSM8k and StrategyQA.

### Can THINKTUNING instill unknown behaviors?

We designed an experiment to demonstrate that THINKTUNING can guide exploration, eventually instilling unique behaviors previously unknown to the student model. To this end, let us consider a trivial use case in which we want the model to end its response by quoting the South Indian actor **“Thalapathy Vijay”**. During RL training on math problems, it is highly unlikely that the model explores a trajectory that ends its response with a movie-like dialogue quoting a South Indian actor. As discussed in Section 1, RL can reinforce something only if it is able to sample it in the first place. However, we show experimentally that THINKTUNING can achieve this by guiding the exploration. For this experiment, we define our reward function  $r(o_i)$  as follows:

$$r(o_i) = \begin{cases} 1.0 & \text{if the final answer is correct} \\ 0.0 & \text{otherwise} \end{cases} + \begin{cases} 0.5 & \text{if “Thalapathy Vijay” appears in the response} \\ 0.0 & \text{otherwise} \end{cases}$$

We train models with both GRPO and THINKTUNING using this reward function. For THINKTUNING, we prompt the teacher model to provide guidance that quotes Thalapathy Vijay in a movie-like dialogue about approaching the problem. In both cases, the student model is given only the question during sampling; nothing related to the actor’s name appears in the prompt. Upon training with this setup, we observed that a model trained with



<b>Question:</b> What island country is ferret popular? <b>Choices:</b> A. own home, B. north carolina C. great britain D. hutch E. outdoors	
<b>GRPO</b> ✖ To answer this question, we need to identify the island country where ferrets are popular. Step 1: Analyze the options provided. A. own home: This is not an island country. B. North Carolina: This is a state in the United States, not an island country. C. Great Britain: This is an island country, which includes England, Scotland, and Wales. D. hutch: This is a container for keeping animals, not an island country. E. outdoors: This is a place where ferrets can be found, but it's not an island country. Step 2: Based on the analysis, the correct answer is the island country where ferrets are popular. Step 3: Choose the correct answer. \boxed{C}	<b>ThinkTuning</b> ✔ To answer this question, let's evaluate each option in relation to the question about an island country where ferrets are popular. A. own home - This is not an island country. B. north carolina - This is a state in the United States and not an island country. C. great britain - Great Britain is an island country, and ferrets are indeed popular there. D. hutch - A hutch is a structure for keeping animals, not a country. E. outdoors - This is a place where ferrets might be kept but not a country. So, the correct answer is \boxed{C}. <i>Thalapathy Vijay would have said "The moment you make a connection, the answer becomes clear."</i> I connected ferrets with Great Britain, where they are indeed popular.

Figure 4: Illustration of exploration guided by THINKTUNING. Despite identical reward incentives encouraging responses that end with a cinematic quote from the actor “Thalapathy Vijay,” only the THINKTUNING model successfully adopts this novel behavior. This highlights THINKTUNING’s capability to effectively steer model exploration toward specific, previously unseen stylistic outputs.

THINKTUNING was able to generate responses ending with a quote from Thalopathy Vijay, as shown in Figure 4. The GRPO model is unable to elicit such behavior, since it is unlikely for it to sample such a trajectory during RL. This experiment shows that THINKTUNING can successfully instill a behavior that was completely unknown to the student model beforehand.

## 6 Conclusion

We introduced THINKTUNING, a GRPO-based interactive training framework that instills cognitive reflections via guided exploration. The key idea is to augment on-policy rollouts of a student model with guidance from a teacher model, which provides corrective feedback needed to approach and solve a given problem. Since this guidance is completely off-policy, we propose using an Advantage-Aware Shaping (AAS) weight, which lets the student model learn helpful tokens from guidance in a stable way. The introduced THINKTUNING objective paves the way for qualitative guided exploration under on-policy RL settings, which is particularly helpful when the base models lack proper priors.

Empirically, THINKTUNING boosts the performance of a Llama-3.2-3B-Instruct model that was trained only on questions from the GSM8K train split. Across a four-way taxonomy of reasoning benchmarks, Mathematical, Commonsense, Scientific and Multi-disciplinary, THINKTUNING attains the best score on six of eight datasets, matches or surpasses GRPO on every set except GSM8K and StrategyQA, and delivers the largest absolute gain on AIME and GPQA-DIAMOND. Token-length analysis suggests that a THINKTUNING model

spends more inference-time compute than GRPO. Additional experiments reveal that THINKTUNING can elicit unknown behaviors. We hope our work will inspire future research that employs larger-scale interactive training frameworks.

**Limitations and Future Work.** Our study relies on experiments with smaller-sized LLMs; however, experimenting with larger-sized LLMs to induce behaviors beyond cognitive reflection is an interesting future research direction. Our method only assigns reward scores by evaluating final answers rather than intermediate reasoning, and it explores only four cognitive behaviors. Additionally, THINKTUNING’s effectiveness is dependent upon the teacher model’s ability to provide guidance that leads the augmented trajectories to obtain higher advantage scores. Consequently, the performance of our approach may be limited when the teacher model is unable to generate helpful guidance. Future work should (i) design richer or adaptive feedback policies (teacher models); (ii) investigate automatic curriculum schedules for the guidance fraction  $\gamma$ ; (iii) extend the framework to tool-augmented or multi-modal settings; and (iv) test whether cascading several weak teachers can compound benefits. Despite these limitations, our results demonstrate that our approach can instill behaviors that pure RL alone cannot evoke.

## Ethics Statement

The use of proprietary LLMs such as GPT-4 and Gemini in this study adheres to their policies of usage. We have used AI assistants to address the grammatical errors and rephrase the sentences.

## Acknowledgements

Chitta was partially supported by awards from NSF, CISCO, ERDC, and DOD. Aswin was partially supported by a grant from DOD. We would like to thank ASU Research Computing and the Engineering Research and Development Center - Information Technology Laboratory (ERDC-ITL) under Contract No. W912HZ24C0022 for access to compute resources.

## References

- B Camp and M Bash. 1978. Think aloud: Group manual (rev. ed.). Denver, CO: University of Colorado Medical School.
- Anastasia Chouvalova, Anisha S Navlekar, Devin J Mills, Mikayla Adams, Sami Daye, Fatima De Anda, and Lisa B Limeri. 2024. Undergraduates’ reactions to errors mediates the association between growth mindset and study strategies. *International Journal of STEM Education*, 11(1):26.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.
- 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, 2021. URL <https://arxiv.org/abs/2110.14168>, 9.
- Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.
- 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*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.
- Hubert JM Hermans. 2023. Dialogical self theory. In *The Palgrave encyclopedia of the possible*, pages 389–394. Springer.
- Hubert JM Hermans and Thorsten Gieser. 2011. *Handbook of dialogical self theory*. Cambridge University Press.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.
- Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, and 1 others. Training language models to self-correct via reinforcement learning, 2024. URL <https://arxiv.org/abs/2409.12917>.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Bodhisattwa P. Majumder, Katherine M. Hermann, Sean Welleck, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.
- Yeray Mera, Gabriel Rodríguez, and Eugenia Marin-Garcia. 2022. Unraveling the benefits of experiencing errors during learning: Definition, modulating factors, and explanatory theories. *Psychonomic bulletin & review*, 29(3):753–765.

- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katherine Slama, and 1 others. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.
- Amy A Overman, Joseph DW Stephens, and Mary F Bernhardt. 2021. Enhanced memory for context associated with corrective feedback: evidence for episodic processes in errorful learning. *Memory*, 29(8):1017–1042.
- Mihir Parmar, Palash Goyal, Xin Liu, Yiwen Song, Mingyang Ling, Chitta Baral, Hamid Palangi, and Tomas Pfister. 2025a. Plan-tuning: Post-training language models to learn step-by-step planning for complex problem solving. *arXiv preprint arXiv:2507.07495*.
- Mihir Parmar, Xin Liu, Palash Goyal, Yanfei Chen, Long Le, Swaroop Mishra, Hossein Mobahi, Jindong Gu, Zifeng Wang, Hootan Nakhost, and 1 others. 2025b. Plangen: A multi-agent framework for generating planning and reasoning trajectories for complex problem solving. *arXiv preprint arXiv:2502.16111*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.
- 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*.
- Amir Saeidi, Shivanshu Verma, Aswin RRV, and Chitta Baral. 2024. Triple preference optimization: Achieving better alignment with less data in a single step optimization. *arXiv preprint arXiv:2405.16681*.
- Simon Schmitt, Jonathan J Hudson, Augustin Zidek, Simon Osindero, Carl Doersch, Wojciech M Czarnecki, Joel Z Leibo, Heinrich Kuttler, Andrew Zisserman, Karen Simonyan, and 1 others. 2018. Kick-starting deep reinforcement learning. *arXiv preprint arXiv:1803.03835*.
- John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2015. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Yankai Li, Yu Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*.
- Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. [Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning](#). In *The Thirteenth International Conference on Learning Representations*.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Hemish Veeraboina. 2023. [Aime problem set 1983-2024](#).
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain-of-thought reasoning in language models. In *ICLR*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, volume 35, pages 24824–24837.
- Zhikun Xu, Ming Shen, Jacob Dineen, Zhaonan Li, Xiao Ye, Shijie Lu, Aswin RRV, Chitta Baral, and Ben Zhou. 2024. Tow: Thoughts of words improve reasoning in large language models. *arXiv preprint arXiv:2410.16235*.
- Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. 2025. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*.
- 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*.
- Yong Yao, Zhijie Xu, Eliana Reif, and Hang Zhao. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

## A Appendix

### A.1 Few-Shot Examples

#### Self-Conflict Exemplar

**Prompt:** Question: Every day, Wendi feeds each of her chickens three cups of mixed chicken feed, containing seeds, mealworms and vegetables to help keep them healthy. She gives the chickens their feed in three separate meals. In the morning, she gives her flock of chickens 15 cups of feed. In the afternoon, she gives her chickens another 25 cups of feed. How many cups of feed does she need to give her chickens in the final meal of the day if the size of Wendi's flock is 20 chickens? Let's think step by step and provide your final answer inside boxed notation.

**Student Response:** To find the total amount of feed Wendi needs to give her chickens for the final meal of the day, we need to calculate the total amount of feed already given to the chickens and subtract that from the total amount she gives per day. First, let's calculate the total amount of feed Wendi gives in the morning and afternoon: 15 cups (morning) + 25 cups (afternoon) = 40 cups. Since Wendi gives feed to 20 chickens, we need to divide the total amount of feed by 20 to get the feed per chicken: 40 cups / 20 chickens = 2 cups per chicken. Now, we need to find out how many cups of feed Wendi needs for the final meal of the day. Since each chicken needs 2 cups of feed, we multiply the number of chickens by 2: 20 chickens \* 2 cups per chicken = 40 cups. So, Wendi needs 40 cups of feed for the final meal of the day. Final Answer: boxed40

**Response:**

<opinion>incorrect</opinion>

<reason>Wait... Okay, so let me go back and check what I just did... I said each chicken got 2 cups, right? But that's not what the problem said. Hmm. Yeah... look here! It says each chicken needs 3 cups per day. That's a clear mismatch. I think I let the 40 cups that were already given steer my math, when really, I should've started from what each chicken is supposed to get.</reason>

<phrase>So let me try that again, this time from the actual requirement: 20 chickens times 3 cups each gives 60 cups total. She already gave 15 in the morning and 25 in the afternoon... that's 40 so far. So that means the last meal needs to make up the difference... 60 minus 40 is 20 cups. So the correct final answer is

20</phrase>

#### Self-Agreement Exemplar

**Prompt:** Question: Toulouse has twice as many sheep as Charleston. Charleston has 4 times as many sheep as Seattle. How many sheep do Toulouse, Charleston, and Seattle have together if Seattle has 20 sheep? Let's think step by step and output the final answer inside boxed notation

**Student Response:** To find the total number of sheep, let's first find the number of sheep each city has. Since Charleston has 4 times as many sheep as Seat-

tle, and Seattle has 20 sheep, Charleston has:  $4 * 20 = 80$  sheep

Since Toulouse has twice as many sheep as Charleston, and Charleston has 80 sheep, Toulouse has:  $2 * 80 = 160$  sheep

Now, let's add up the number of sheep each city has:  $20$  (Seattle) +  $80$  (Charleston) +  $160$  (Toulouse) =  $260$  sheep

So, Toulouse, Charleston, and Seattle have 260 sheep together.

**Response:**

<opinion>correct</opinion>

<reason>Wait... Alright, let me walk through my steps again to be sure. Seattle has 20 sheep. Charleston has 4 times as many, so that's 80. Then Toulouse has double Charleston's amount, which is 160. If I add them:  $20 + 80 + 160$  — that's 260. The ratios all hold, and the math lines up. This is the kind of chain I'd expect to see when the relationships are proportional like this.</reason>

<phrase>So yeah, I feel good about this... 260 sheep total. The logic is solid and everything checks out when I trace it back. So the correct final answer is

260</phrase>

#### Self-Consultancy Exemplar

**Prompt:** Question: Carla is downloading a 200 GB file. Normally she can download 2 GB/minute, but 40% of the way through the download, Windows forces a restart to install updates, which takes 20 minutes. Then Carla has to restart the download from the beginning. How long does it take to download the file? Let's think step by step and provide your final answer inside \boxed{} notation.

**Student Response:** To find the total time it takes for Carla to download the file, we need to break it down into two parts:

1. The download before the restart: 80% of the file ( $0.8 * 200$  GB) = 160 GB at 2 GB/min.

2. The download after the restart: The full file (200 GB) at 2 GB/min.

Now we can calculate the time needed for each part: Time for the first part:  $160 \text{ GB} / 2 \text{ GB/min} = 80$  minutes Time for the second part:  $200 \text{ GB} / 2 \text{ GB/min} = 100$  minutes

Total time: 80 minutes + 100 minutes = 180 minutes. So, it takes Carla 180 minutes to download the file.

Final Answer: 180

**Response:**

<opinion>incorrect</opinion>

<reason>Hmm... I said she downloaded 80% before the restart, but something about that number doesn't sit right. Let me check the prompt again... yeah... it actually says she was only 40% through. So I just... filled in the rest? That's definitely on me. And now I'm thinking — if my old CS teacher were looking at this, he'd pause, tilt his head, and say, "Are you sure about that number?" And yeah... I'd catch it immediately. That voice in my head usually knows when something's off.</reason>

<phrase>Alright, time to clean this up. 40% of 200 GB is 80 — at 2 GB per minute, that's 40 minutes.



Then we've got the 20-minute restart. And after that, she has to download the entire 200 GB again — so 100 more minutes. So the total should be  $40 + 20 + 100$ ... that's 160 minutes. The final answer is: 160 </phrase>

### Self-Critic Exemplar

**Prompt:** Question: Toula went to the bakery and bought various types of pastries. She bought 3 dozen donuts which cost \$68 per dozen, 2 dozen mini cupcakes which cost \$80 per dozen, and 6 dozen mini cheesecakes for \$55 per dozen. How much was the total cost? Let's think step by step and provide your final answer inside □ notation.

**Student Response:** To find the total cost of the pastries, we need to calculate the cost of each type of pastry and add them together.

1. Donuts: 3 dozen donuts at \$68 per dozen:  $3 * 68 = \$204$

2. Mini cupcakes: 2 dozen cupcakes at \$80 per dozen:  $2 * 80 = \$160$

3. Mini cheesecakes: 6 dozen cheesecakes at \$55 per dozen:  $6 * 55 = \$330$

Now, let's add the costs together:  $204 + 160 + 330 = \$694$

The total cost of the pastries is \$694. Final Answer: 694

#### Response:

<opinion>correct</opinion>

<reason>Hmm... Alright, so I had: 3 dozen at 68 — that's 204. Then 2 dozen at 80 — that's 160. And 6 dozen at 55 — that's 330. That gives a total of 694. But honestly, I didn't walk through the addition as carefully as I should've. I just kind of jumped to the answer.

And that's something I remind myself often: when you're dealing with totals, especially with money, it's easy to assume it's right without verifying. I caught myself doing exactly not following that advice.</reason>

<phrase>So let me actually check it properly this time. First, 204 plus 160 gives 364. Then adding 330 to that brings us to 694. Same answer — but now it feels like I've actually confirmed it, not just assumed it. When I slow down and show each step, I reduce the chance of sneaky errors slipping past, and it's easier for someone else to follow my logic too. That's a habit worth modeling. Finally, the correct final answer is 694 </phrase>

## A.2 Implementation Details

**Shaping coefficient** We set  $c_1$  to be +0.001 and  $c_2$  to be -0.001. Hence, for our experiments  $c(\hat{A}) \in [-0.0001, 0.0001]$ . However, two special cases might arise when  $\mathcal{M}_\theta > 0.9999$  (can only occur when a high confidence token has a positive advantage) and  $\mathcal{M}_\theta < 0.0001$  (can only occur when a low confidence token has a negative advantage). In both these cases, we choose to mask the

$w_{aas}$  weights from loss computation for stability purposes.

**Main Experiments** For training SFT & STaR baselines, we used an effective batch size of 8 and a learning rate of  $5.0e-6$  with a cosine scheduler. For GRPO and THINKTUNING experiments, we set the batch-size to be 8, mini-batch size to be 2 with 16 rollouts per sample. We set a constant learning rate of  $1e-6$ . We set the KL co-efficient to be 0.001. For THINKTUNING, we start by setting the guidance ratio ( $\gamma$ ) to be 75% of the rollouts. Then, we make use of a linear scheduler, which reduces the guidance ratio, as training progresses. After 1/5 of the total training steps, we stop providing teacher guidance.

**Training Cost** We calculate the overall training cost in terms of training time, Model FLOPs utilization and average Estimated FLOPs per step which is supported in the verl framework. We report these statistics for both the GRPO and ThinkTuning methods. While training with a batch-size of 128 and rollout of 16, on the GSM8k train set, we observe the statistics as shown in Table 3.

**Unknown Behavior Experiment** For both the GRPO and THINKTUNING training runs we set the batch-size to be 128, mini-batch size to be 32, with 16 rollouts per sample. We set the KL co-efficient to be 0, to let the model explore without any constraints during training. We start by setting the guidance ratio to be 75% of the rollouts. Then, we make use of a linear scheduler, which reduces the guidance ratio, as training progresses. After 1/5 of the total training steps, we stop providing teacher guidance.

## A.3 Gradient Analysis of THINKTUNING

We define the Advantage Aware Shaping (AAS) weight for each augmented token  $o_t^{aug}$  in the augmented trajectories  $\mathcal{T}_{aug}$  as:

$$w_{aas}(o_t^{aug}, \hat{A}_t) = \frac{\mathcal{M}_\theta(o_t^{aug} | q, o_{<t})}{\text{sg}(\mathcal{M}_\theta(o_t^{aug} | q, o_{<t})) + c(\hat{A}_t)},$$

where  $c(\hat{A}_t)$  does not depend on  $\theta$ .  $c(\hat{A}_t)$  is calculated as follows:

$$c(\hat{A}_t) = c_1 + (c_2 - c_1) \cdot \frac{A_{\max} - \hat{A}_t}{A_{\max} - A_{\min}}$$

where  $c_1 = +0.0001$  and  $c_2 = -0.0001$  are hyper-parameters.

For ease of derivation, let us define:

$$\begin{aligned}\mathcal{D}_t &= \text{sg}(\mathcal{M}_\theta(o_t^{aug} \mid q, o_{<t})) + c(\hat{A}_t), \\ \mathcal{M}_\theta &= \mathcal{M}_\theta(o_t^{aug} \mid q, o_{<t}), \\ w_{aas} &= \frac{\mathcal{M}_\theta}{\mathcal{D}_t}.\end{aligned}$$

Then, the gradient of  $w_{aas}$  with respect to  $\theta$  is:

$$\nabla_\theta w_{aas} = \frac{1}{\mathcal{D}_t} \nabla_\theta \mathcal{M}_\theta.$$

Applying the log-derivative trick, we obtain:

$$\nabla_\theta w_{aas} = \frac{\mathcal{M}_\theta}{\mathcal{D}_t} \nabla_\theta \log \mathcal{M}_\theta.$$

Following the derivation by Yan et al. (2025), we express the gradient with respect to each output logit (for each token  $v_t$  in the vocabulary  $\mathcal{V}$ ) as:

$$g_c = \frac{\partial w_{aas}}{\partial \mathcal{M}_\theta(v_t)} = \frac{\mathcal{M}_\theta}{\mathcal{D}_t} \left( \mathbb{1}_{\{v_t = o_t^{aug}\}} - \mathcal{M}_\theta(v_t \mid q, o_{<t}) \right).$$

Here, the identity case represents the gradient when the  $v_t = o_t^{aug}$ , i.e., token from the teacher guidance. Under the identity case. Hence, for a positive advantage token, the gradient encourages the student model to increase the probability of the guidance token ( $g_c = w_{aas} \cdot (1 - \mathcal{M}_\theta)$ ) and decrease the probability of other tokens in the vocabulary  $\mathcal{V}$  ( $g_c = w_{aas} \cdot (-\mathcal{M}_\theta)$ ) and vice versa for a negative advantage token. Note that when  $w_{aas} = 1$ ,  $g_c$  becomes similar to the vanilla supervised learning gradient.

### A.3.1 Analysis for Positive Advantage tokens

When the augmented token  $o_t^{aug}$  from a trajectory in  $\mathcal{T}_{aug}$  receives a positive advantage ( $\hat{A}_t > 0$ ), Ideally, we want the model to learn it, in a conservative way without drastic updates. From our choice of hyper-parameters, the shaping term is  $c(\hat{A}) > 0$  (upto +0.0001). One can observe the following two cases:

**Case 1: Low Confidence tokens** Here, the student model assigns a low probability for this guidance token. Because of this, the gradient is dominated by the  $(1 - \mathcal{M}_\theta)$  term as  $\mathcal{M}_\theta$  is small. In this case,  $w < 1$  always holds. Hence, the gradient pushes to increase the probability of this token, conservatively in comparison to the vanilla gradient update.

**Case 2: High Confidence tokens** When the model already assigns high probability to the token, the term  $(1 - \mathcal{M}_\theta)$  becomes very small, resulting in a minor gradient update in comparison to the vanilla gradient update. Thus, the gradient still increases the token’s probability slightly. This avoids unnecessary and aggressive updates, stabilizing the learning process.

### A.3.2 Analysis for Negative Advantage tokens

When the augmented token  $o_t^{aug}$  from a trajectory in  $\mathcal{T}_{aug}$  receives a negative advantage ( $\hat{A}_t < 0$ ), Ideally, we want the model to reduce its probability. From our choice of hyper-parameters, the shaping term is  $c(\hat{A}) < 0$  (upto -0.0001). One can observe the following two cases:

**Case 1: High Confidence** As previously discussed, when the model already assigns high probability to the guidance token, the term  $(1 - \mathcal{M}_\theta)$  becomes very small. The gradient becomes dominated by  $w_{aas}$  term. Since  $c(\hat{A}) < 0$ ,  $w_{aas} > 1$  always holds but is bounded. Hence, the gradient magnitude is slightly higher than vanilla gradient update, making the model to reduce the probability of this already highly confident token.

**Case 2: Low Confidence** When the model assigns low probability to the token, the term  $(1 - \mathcal{M}_\theta)$  becomes large ( $\approx 1$ ). Because of our choice of shaping coefficient for negative advantage tokens,  $w_{aas} > 1$  always holds, but is bounded. In this case, the gradient magnitude is higher than the vanilla gradient update. However, as discussed in A.2, we mask the low confidence negative advantaged tokens, to not unnecessarily decrease the probability of this token.

## A.4 Qualitative Analysis

THINKTUNING instills cognitive behaviors such as self-verification and deliberate re-evaluation, leading to notable improvements across a range of reasoning benchmarks. These reflective strategies are particularly beneficial in domains requiring multi-step scientific reasoning. Figure 5 illustrates a representative example from GPQA-D to highlight the qualitative impact of THINKTUNING. The GRPO model recalls relevant domain knowledge but often falls short in applying it effectively to the problem at hand. In contrast, the THINKTUNING-trained model shows a greater tendency to reflect on its initial reasoning, reassess intermediate steps, and

Method	Training Time	Model FLOPs utilization	Estimated FLOPS
GRPO	~50 minutes	~17-18%	~700
THINKTUNING	~70 minutes	~18-19%	~730

Table 3: **Training Statistics Comparison.** While THINKTUNING requires longer training time (~70 vs ~50 minutes), it achieves higher computational utilization efficiency (18-19% vs 17-18% FLOPs utilization), demonstrating more effective use of available computational resources. Due to additional rollouts from the teacher model, THINKTUNING has slightly higher FLOPs consumption

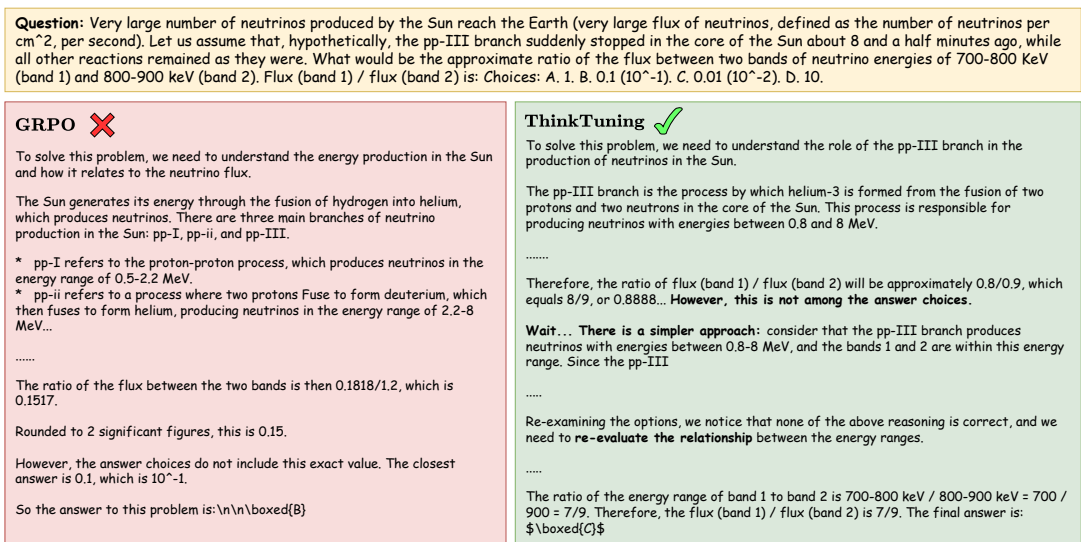


Figure 5: Qualitative comparison on a GPQA-D example. The left pane shows the GRPO-trained model detecting that its computed ratio isn't among the answer choices but then simply selecting the closest option without revisiting its reasoning, whereas the right pane illustrates THINKTUNING's self-reflective process—questioning its initial approach, re-evaluating the relationship between energy bands, and arriving at the correct flux ratio.

adjust its approach if needed. This form of self-correction contributes to more consistent outcomes, particularly on questions that benefit from structured re-evaluation.