ThinkTuning: Instilling Cognitive Reflections without Distillation

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

The 2025 Conference on Empirical Methods in Natural Language Processing
Suzhou, China - November 4–9, 2025

Introduction

Introduction

Large Language Models (LLMs) have significantly transformed the AI landscape, by achieving human-like performance on various downstream tasks.

Model Parameters ↑ ⟶ Downstream Performance ↑

More recently,

Inference time/compute ↑ ⟶ Downstream Performance ↑

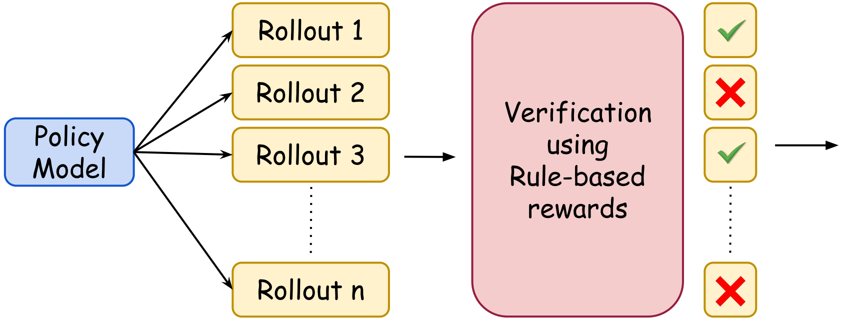
Models learn to produce ***long reasoning chains***, with sophisticated behaviors like ***self-reflection***, ***self-correction***, and ***multi-step reasoning***.

How?

Reinforcement **L**earning with **V**erifiable **R**ewards **(RLVR)!!!**

Introduction

Reinforcement **L**earning with **V**erifiable **R**ewards **(RLVR) -** GRPO variant



$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}\Big[q \sim \mathcal{D},\ \{o_i\}_{i=1}^{n} \sim \mathcal{M}_{\theta_{\mathrm{old}}}(O \mid q)\Big] \Bigg\{ \frac{1}{n}\sum_{i=1}^{n} \frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \min\Big[\frac{\mathcal{M}_{\theta}(o_{i,t}|q,o_{i,<t})}{\mathcal{M}_{\theta_{\mathrm{old}}}(o_{i,t}|q,o_{i,<t})} \hat{A}_{i,t},\ \mathrm{clip}\Big(\frac{\mathcal{M}_{\theta}(o_{i,t}|q,o_{i,<t})}{\mathcal{M}_{\theta_{\mathrm{old}}}(o_{i,t}|q,o_{i,<t})}, 1-\epsilon,\ 1+\epsilon\Big)\, \hat{A}_{i,t}\Big] - \beta\, D_{KL}\big[\mathcal{M}_{\theta} \,\|\, \mathcal{M}_{\mathrm{ref}}\big] \Bigg\}$$

However, there is a constraint!

RL can reinforce something only if it is able to sample it in the first place

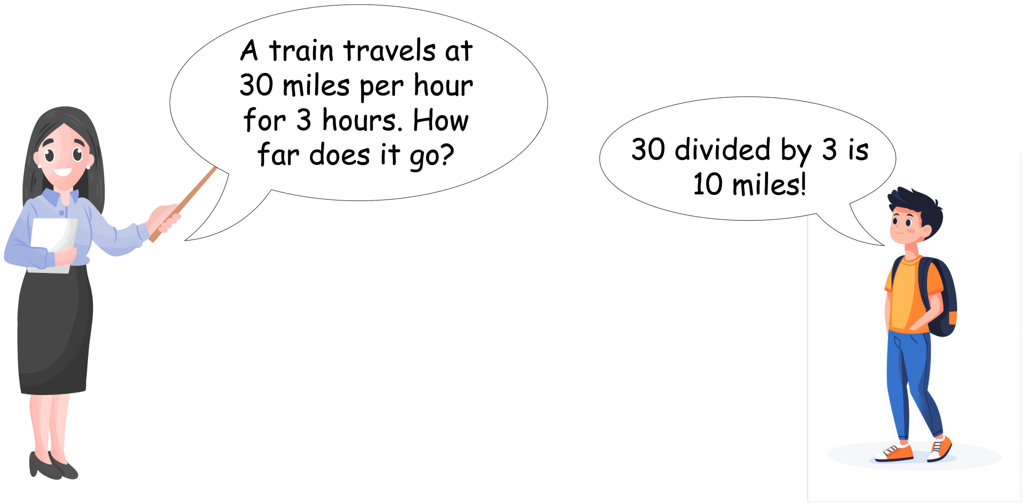
Indeed, a recent study [1] shows that RL applied on the Llama 3.2-family of models struggles to elicit sophisticated thinking behaviors, in comparison to Qwen models.

[1] 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.

Introduction

How can we train the models that don't exhibit such thinking behavior to develop it in the first place?

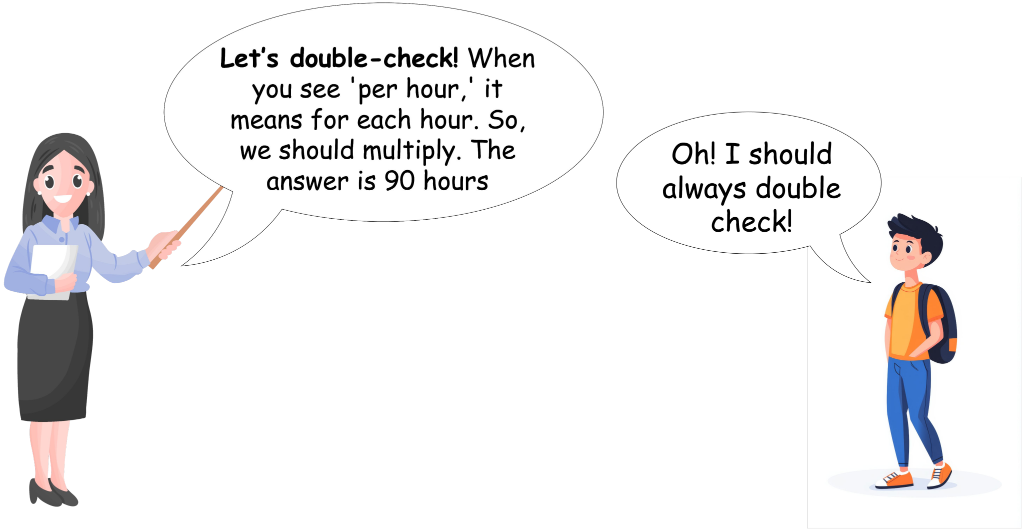
We propose a simple idea inspired from classroom practice!



Introduction

How can we train the models that don't exhibit such thinking behavior to develop it in the first place?

We propose a simple idea inspired from classroom practice!

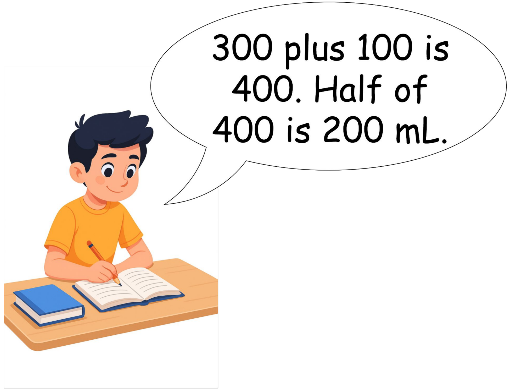


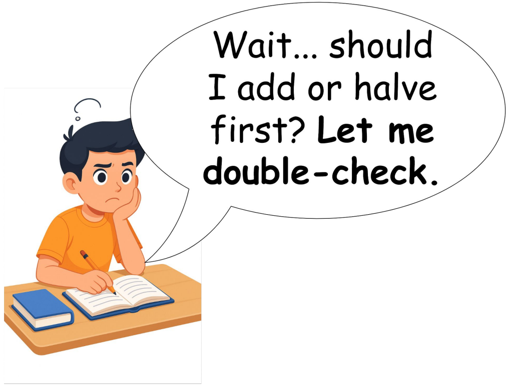
Introduction

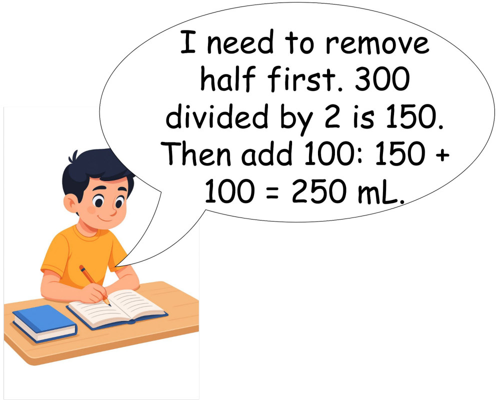
How can we train the models that don't exhibit such thinking behavior to develop it in the first place?

We propose a simple idea inspired from classroom practice!

HomeWork: A tank contains 300 mL of solution. You remove half of it, then add 100 mL of water. How much liquid is in the tank now?

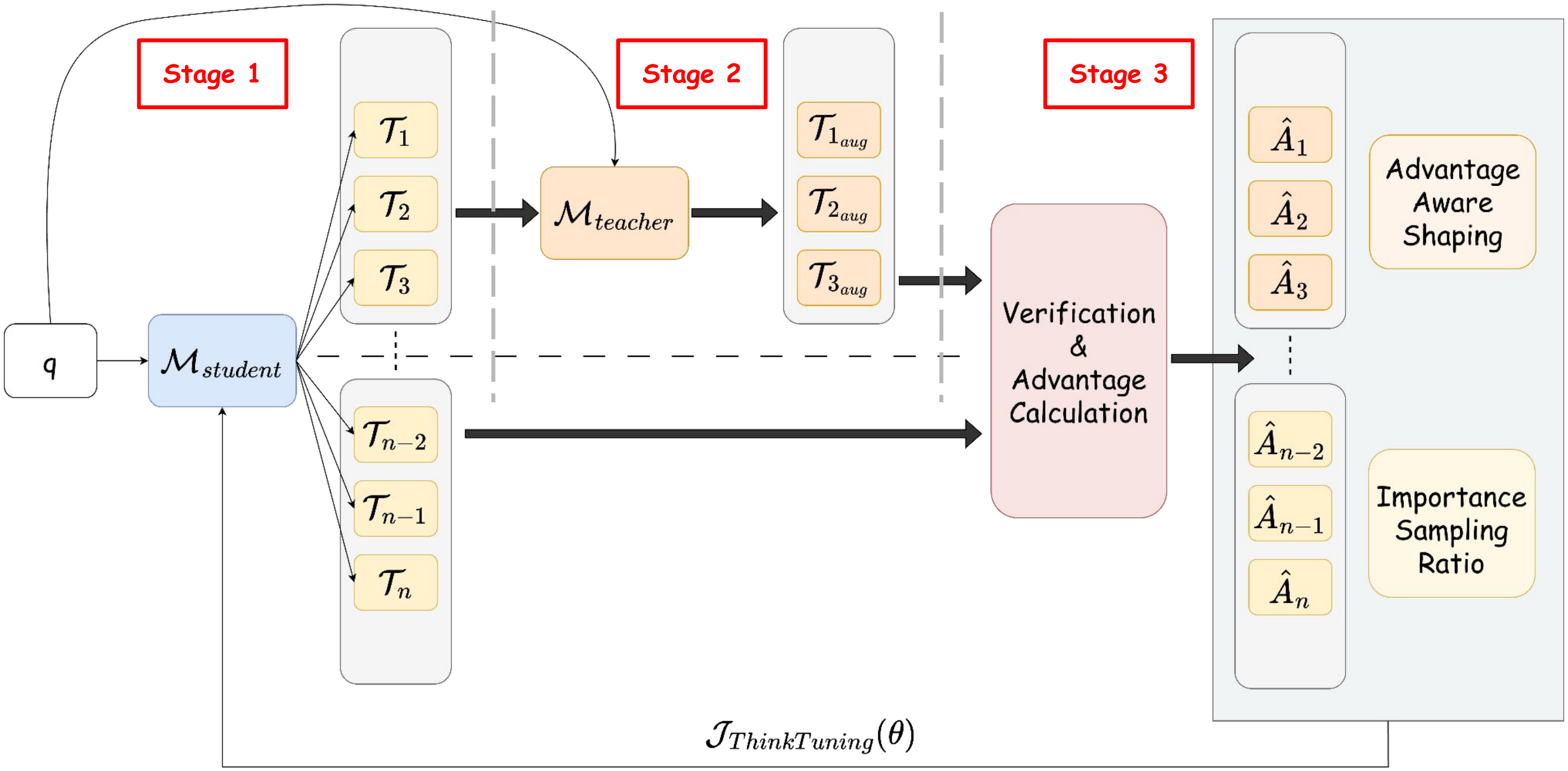






Method

Framework

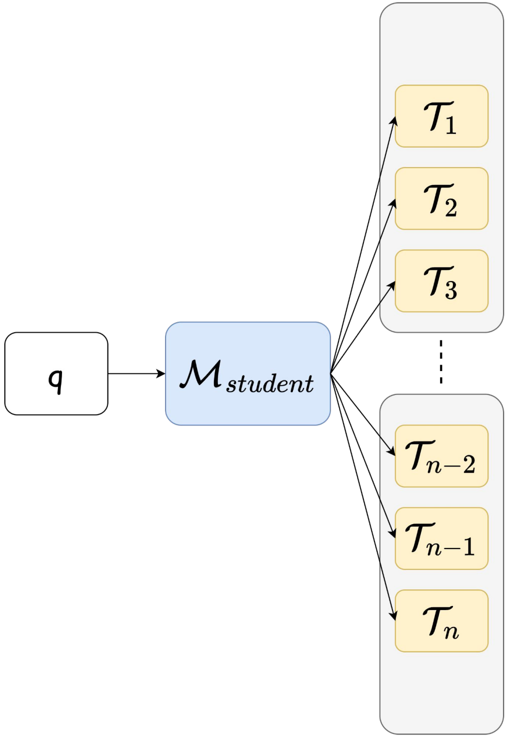


Stage 1
Stage 2
Stage 3
q
$\mathcal{M}_{student}$
$\mathcal{T}_1$
$\mathcal{T}_2$
$\mathcal{T}_3$
$\mathcal{T}_{n-2}$
$\mathcal{T}_{n-1}$
$\mathcal{T}_n$
$\mathcal{M}_{teacher}$
$\mathcal{T}_{1_{aug}}$
$\mathcal{T}_{2_{aug}}$
$\mathcal{T}_{3_{aug}}$
Verification & Advantage Calculation
$\hat{A}_1$
$\hat{A}_2$
$\hat{A}_3$
$\hat{A}_{n-2}$
$\hat{A}_{n-1}$
$\hat{A}_n$
Advantage Aware Shaping
Importance Sampling Ratio
$\mathcal{J}_{ThinkTuning}(\theta)$

Stage 1: Student Responses

For a given question **q**, we sample **n** responses.

We sample responses at a **temperature of 1.0** to observe diversity.



Stage 2: Teacher Guidance

We select γ number of trajectories and pass it to a **teacher model** for guidance.

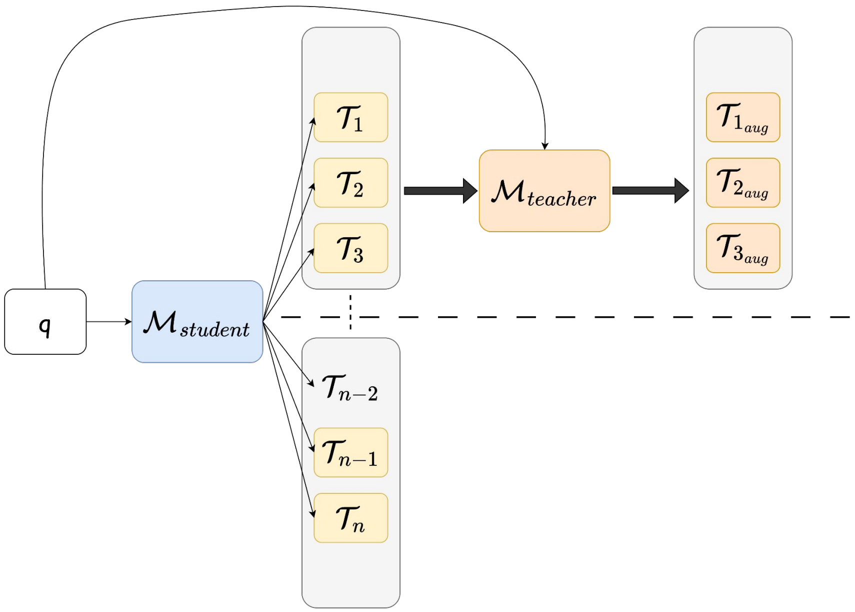
Guidance Mechanism:
We give few-shot examples of four self-reflective cognitive behaviors (first person perspective):

Self-Conflict: challenging one's own response by presenting conflicting perspectives.

Self-Critique: identifying weaknesses in their response and suggesting improvements.

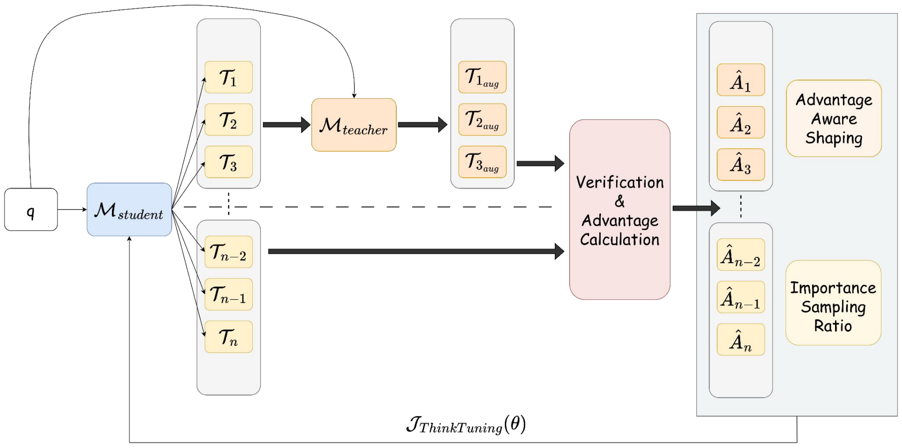
Self-Agreement: affirming and justifying the strengths in their response.

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.



Stage 3: Student Training

The feedback obtained from the teacher model **is augmented** to the student rollouts.



<u>Off-Policy Guidance Tokens</u>

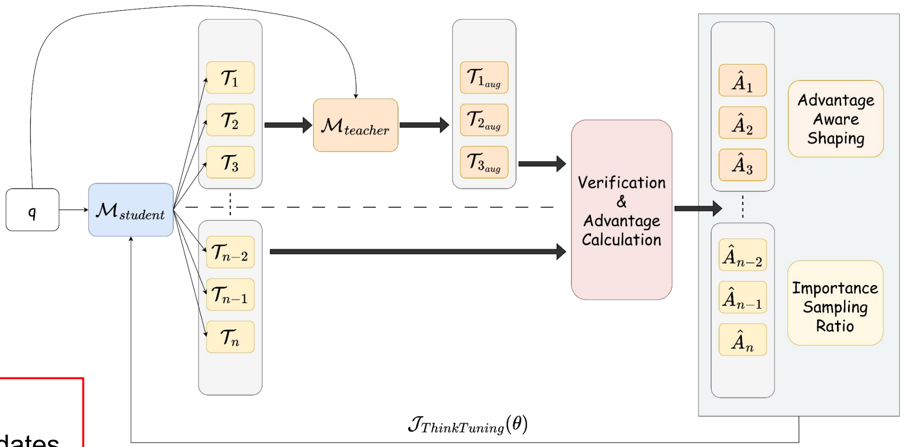
q → $\mathcal{M}_{student}$ → ..... Hence the final answer is X </s> → $\mathcal{M}_{teacher}$ → ..... Hence the final answer is X. Hmm... Wait... Let me recheck my answer ....

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

$$\boxed{\frac{\mathcal{M}_\theta(\text{"Hmm"}|q,r_1)}{\mathcal{M}_{\theta_{\text{old}}}(\text{"Hmm"}|q,r_1)}}$$

During Off-Policy updates, importance sampling weight can be very high, leading to massive clipping or gradients (if not clipped)

Stage 3: Student Training



We propose Advantage-Aware-Shaping (AAS) weight for off-policy tokens

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

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

$$c(\hat{A}) \in [-0.0001, 0.0001]$$

Serves as **a knob** that enables us to **control the magnitude** of gradient updates for the **off-policy teacher tokens**.

Gradient Analysis (for chosen c values)

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

High Confidence Tokens

When A > 0, the gradient increases the token's probability slightly, while preventing overfitting to the already highly confident token.

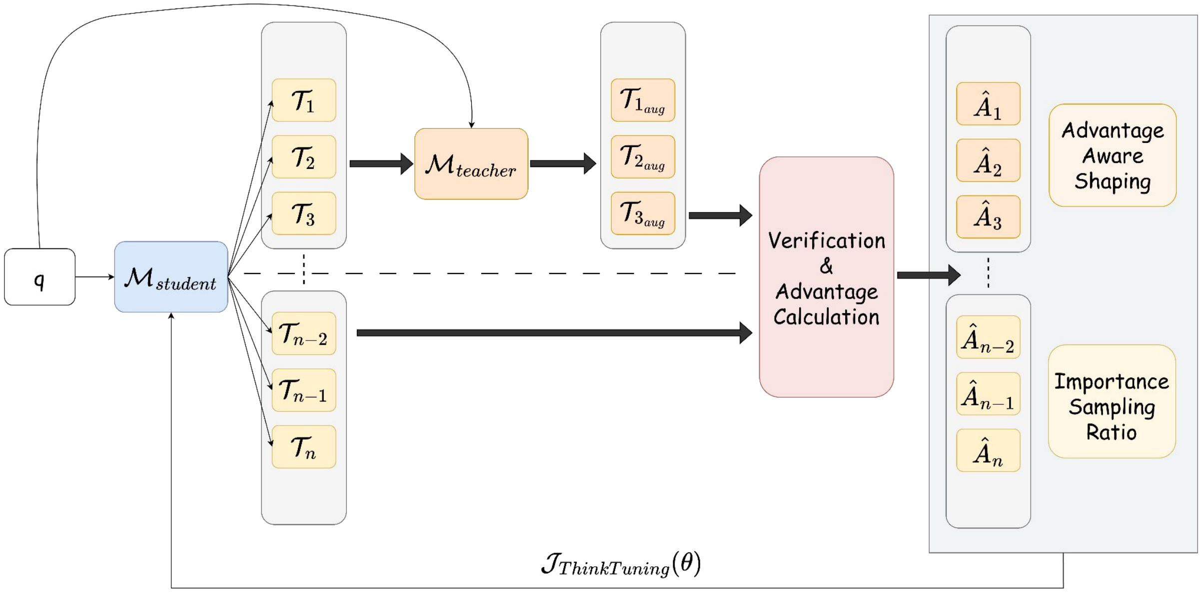
When A < 0, the gradient magnitude is slightly higher than vanilla gradient update, making the model to reduce the probability of this already highly confident token.

Low Confidence Tokens

When A > 0, the gradient pushes to increase the probability of this token, conservatively in comparison to the vanilla gradient update.

When A < 0, beyond a confidence threshold, we mask the low confidence negative advantaged tokens, to not unnecessarily decrease the probability of this token.

Overall Algorithm



Algorithm 1 THINKTUNING

Input: Initial Student model $\mathcal{M}_{student_{\theta_{init}}}$, Teacher model $\mathcal{M}_{teacher}$, guidance fraction γ, hyperparameter set $(\epsilon, \beta, c_1, c_2, k)$

$\mathcal{M}_{student_\theta} \leftarrow \mathcal{M}_{student_{\theta_{init}}}$

for training step=1 **to** I **do**
  $\mathcal{M}_{student_{old}} \leftarrow \mathcal{M}_{student_\theta}$
  Sample mini-batch $\mathcal{D}_b \subset \mathcal{D}$

  // Student acts & Teacher helps
  for all q $\in \mathcal{D}_b$ **do**
    if training step $\leq k$ **then**
      $\{o\}_{i=1}^{n} \sim$ Guide$(q, \mathcal{M}_{student_{old}}, \mathcal{M}_{teacher}, \gamma)$
    else
      $\{o\}_{i=1}^{n} \sim \mathcal{M}_{student_{old}}(O \mid q)$
    end if
  end for

  // Reward calculation and Advantage estimation
  Compute the rewards $r_i = r(o_i)$ for each response
  Compute group-normalized advantage $\hat{A}_{i,t}$ for all tokens

  for mini-batch step = 1 **to** μ **do**
    if training step $\leq k$ and $o_i \in \mathcal{T}_{\text{aug}}$ **then**
      Calculate $w_{\text{aas}}(o_i)$
    else
      Calculate $w(o_i)$
    end if
    $\mathcal{M}_{student_\theta} \leftarrow \text{argmax}_\theta\ \mathcal{J}_{\text{THINKTUNING}}(\theta)$
  end for
end for
Output: Final think-tuned model $\mathcal{M}_{student_\theta}$

Experiment & Results

Experimental Setup

Baselines

Prompt-based Methods
Zero-Shot
Self-Verify
Self-Correct
S1-Budgeting (Token Budget = 2048)

Training-based Methods
SFT
STaR

RL Baseline
GRPO

Model

Llama-3.2-3B-Instruct

Benchmarks

GSM8k
Math-500
AIME
CommonSenseQA
ARC-Challenge
GPQA-Diamond
StrategyQA
MMLU-Pro

Main Results

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	$51.30_{\pm 0.42}$	$25.72_{\pm 0.54}$	$9.01_{\pm 0.31}$	$54.21_{\pm 0.44}$	$59.51_{\pm 0.27}$	$\underline{26.57}_{\pm 0.99}$	$57.88_{\pm 0.80}$	$28.59_{\pm 0.10}$
SFT	$62.27_{\pm 0.61}$	$29.00_{\pm 0.49}$	$6.07_{\pm 0.43}$	$65.91_{\pm 0.24}$	$70.90_{\pm 0.71}$	$24.49_{\pm 0.82}$	$64.12_{\pm 0.65}$	$\underline{36.07}_{\pm 0.07}$
STaR	$73.54_{\pm 0.22}$	$40.78_{\pm 0.35}$	$8.91_{\pm 0.29}$	$67.91_{\pm 0.30}$	$77.24_{\pm 0.21}$	$21.46_{\pm 0.86}$	$\underline{66.84}_{\pm 0.41}$	$34.69_{\pm 0.12}$
GRPO	$\mathbf{78.89}_{\pm 0.84}$	$\underline{45.46}_{\pm 1.55}$	$\underline{12.03}_{\pm 0.33}$	$\underline{69.86}_{\pm 0.52}$	$\underline{79.13}_{\pm 0.21}$	$24.19_{\pm 0.75}$	$\mathbf{70.68}_{\pm 0.35}$	$\underline{36.07}_{\pm 0.07}$
THINKTUNING	$\underline{74.22}_{\pm 0.13}$	$\mathbf{47.54}_{\pm 0.46}$	$\mathbf{14.26}_{\pm 0.38}$	$\mathbf{70.43}_{\pm 0.19}$	$\mathbf{79.80}_{\pm 0.24}$	$\mathbf{28.18}_{\pm 0.63}$	$66.52_{\pm 0.41}$	$\mathbf{37.21}_{\pm 0.11}$

Comparison with Prompt-based methods: ThinkTuning outperforms all prompting-based baselines, achieving up to +9.4% accuracy gains across diverse reasoning benchmarks.

Comparison with Training-Based methods: SFT harms reasoning, reducing accuracy by up to 8% even on in-domain GSM8k. While STaR recovers some performance over SFT, our method surpasses it across all benchmarks with up to +6.8% gains with respect to zero-shot baseline.

Comparison with GRPO: Our method 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

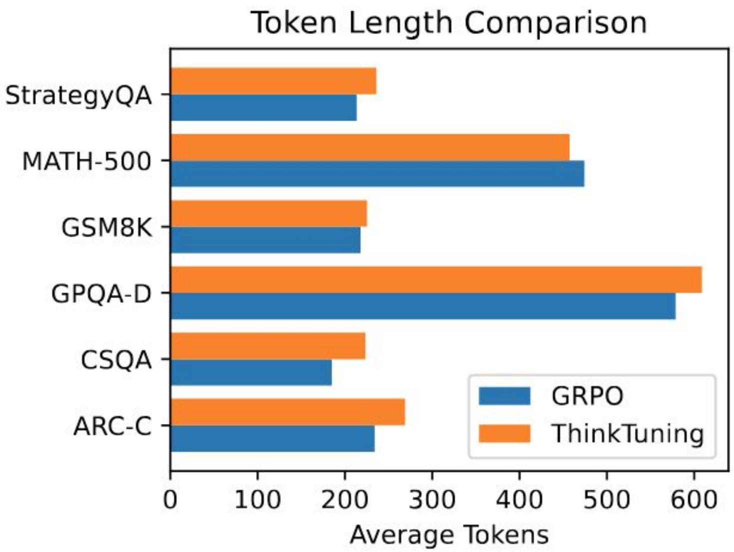
Analysis

Does *ThinkTuning* scale inference time?

ThinkTuning model generates longer responses (3.4–20.8% more tokens) across benchmarks, indicating deeper reasoning.

Increased token usage reflects boosted performance on complex tasks like GPQA-D and MATH-500.

In MATH-500, ThinkTuning model outperforms GRPO even with fewer tokens (−3.6%), showing more efficient reasoning.



Can *ThinkTuning* instill unknown behaviors?

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 movielike dialogue quoting **a South Indian actor**. Note that, **we do not give** anything instructing the model to respond this particular way during sampling.

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

Question: What island country is ferret popular?
Choices: A. own home, B. north carolina C. great britain D. hutch E. outdoors

GRPO ✗

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}

ThinkTuning ✓

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}. **Thalapathy Vijay would have said "The moment you make a connection, the answer becomes clear."** I connected ferrets with Great Britain, where they are indeed popular.

Conclusion

Our Contributions

- We introduced **ThinkTuning**, a GRPO-based interactive training framework that **instills cognitive reflections via guided exploration**.

- Empirically, ThinkTuning boosts the performance of a Llama-3.2-3B-Instruct model that was trained only on questions from the GSM8K train split.

- Additional experiments reveal that ThinkTuning **can elicit unknown behaviors**, by **steering/guiding the exploration** on-the-fly during RL.

- We hope our work will inspire future research that employs **large-scale interactive training frameworks**.

End