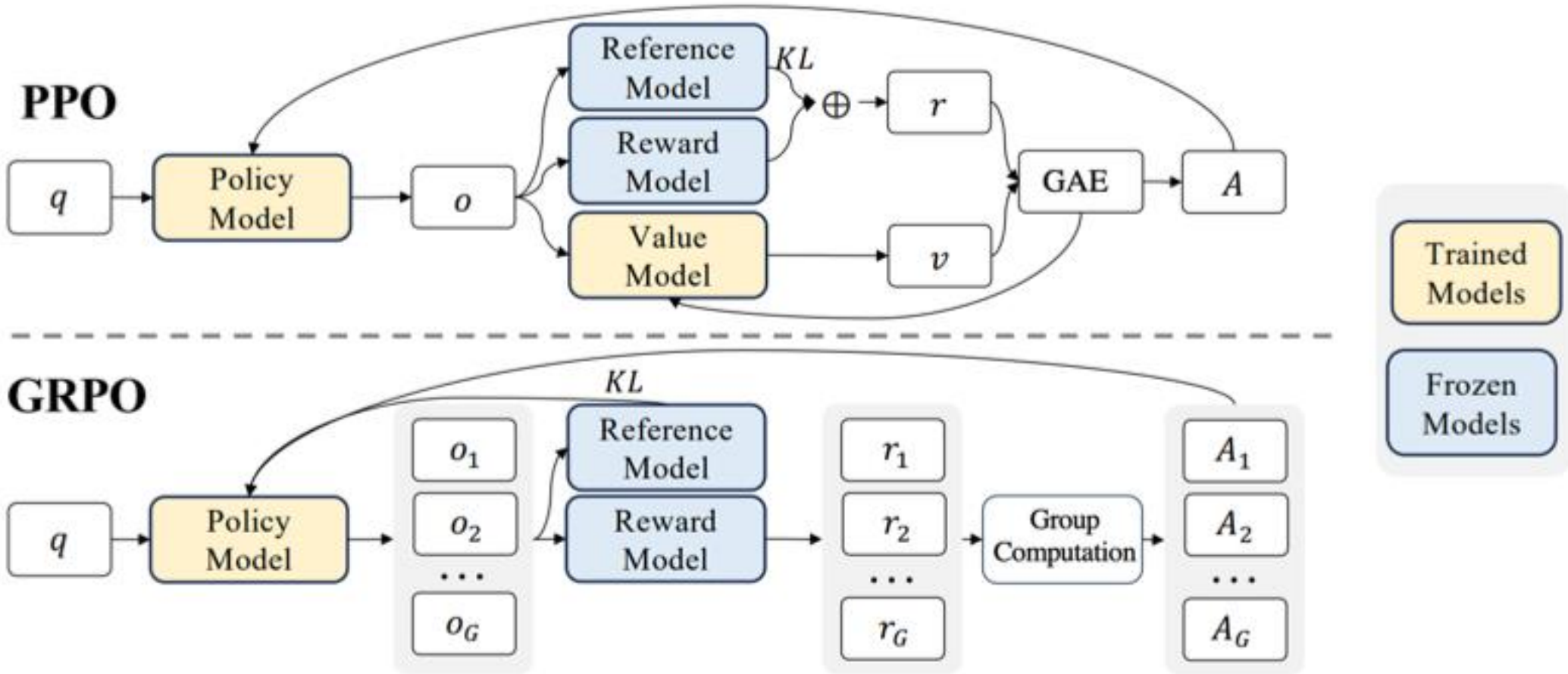
What does RL improve for Visual Reasoning? A Frankenstein-Style Analysis

Xirui Li[1,*], **Ming Li**[1,*], **Tianyi Zhou**[2]

[1]University of Maryland, [2]Mohamed bin Zayed University of Artificial Intelligence
[*]Co-first Author

Background

RLVR example:

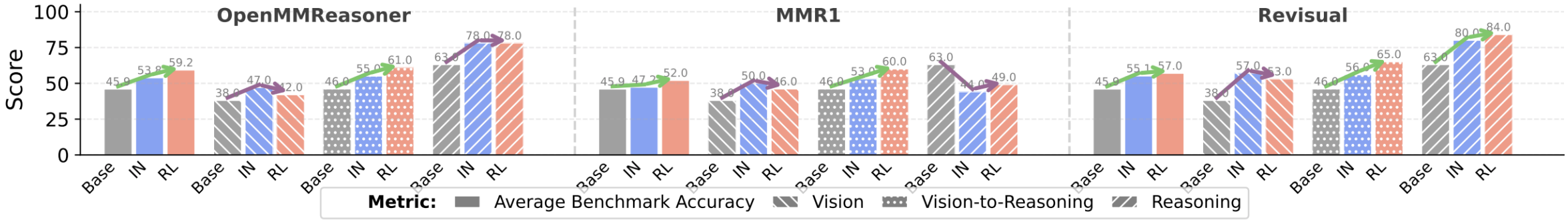


Motivation: Inconsistent Fine-grained Improvements

Fine-grained evaluation metrics targeting vision, vision-to-reasoning alignment, and pure reasoning ability.

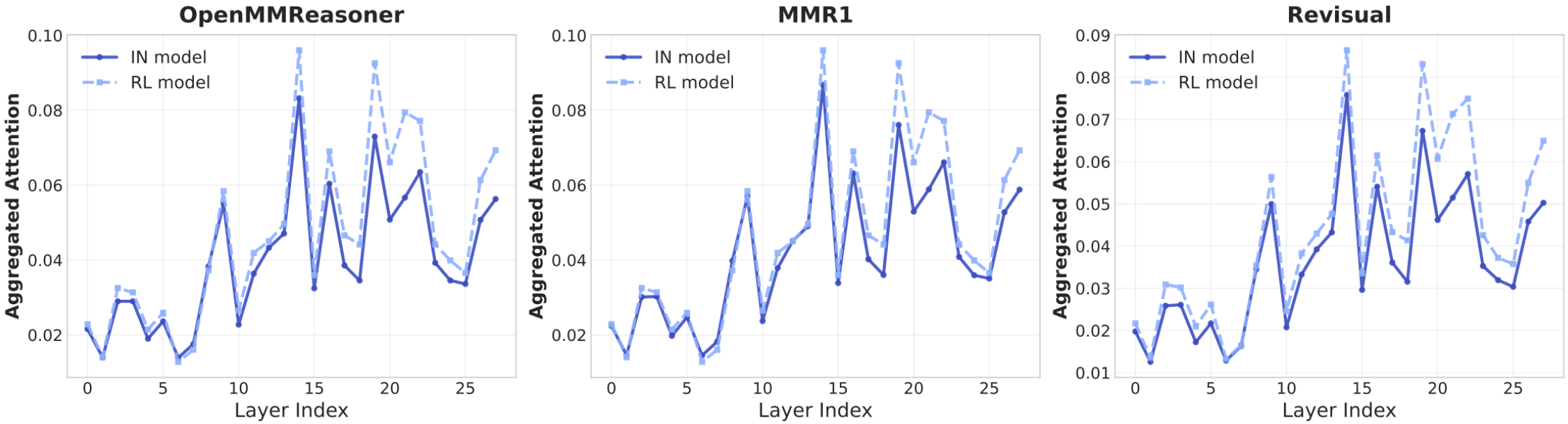
Metric	Task	Definition
M_{vis}	General VQA	$\frac{1}{N}\sum_{n=1}^{N}\mathbb{I}\big[f(i_n,p_n)=y_n \wedge f(b_n,p_n)\neq y_n\big]$
M_{v2r}	Math VQA	$\frac{1}{N}\sum_{n=1}^{N}\mathbb{I}\big[f(i_n,p_n)=y_n \wedge f(b_n,d_n,p_n)=y_n\big]$
M_{rea}	Textual Math	$\frac{1}{N}\sum_{n=1}^{N}\mathbb{I}\big[f(p_n)=y_n\big]$

Average Benchmark Accuracy versus Fine-Grained Abilities (Vision, Reasoning, and Vision-to-Reasoning Alignment)



Motivation: Consistent Attention Patterns

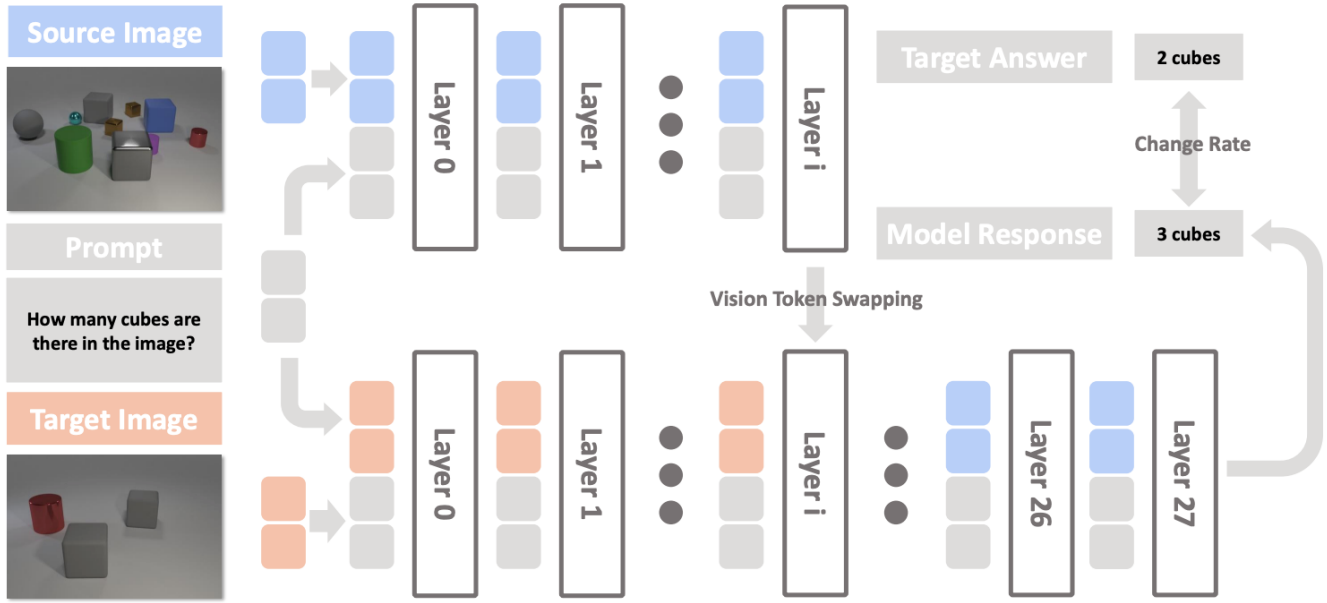
Aggregated Attention from Reasoning Tokens to Vision Tokens

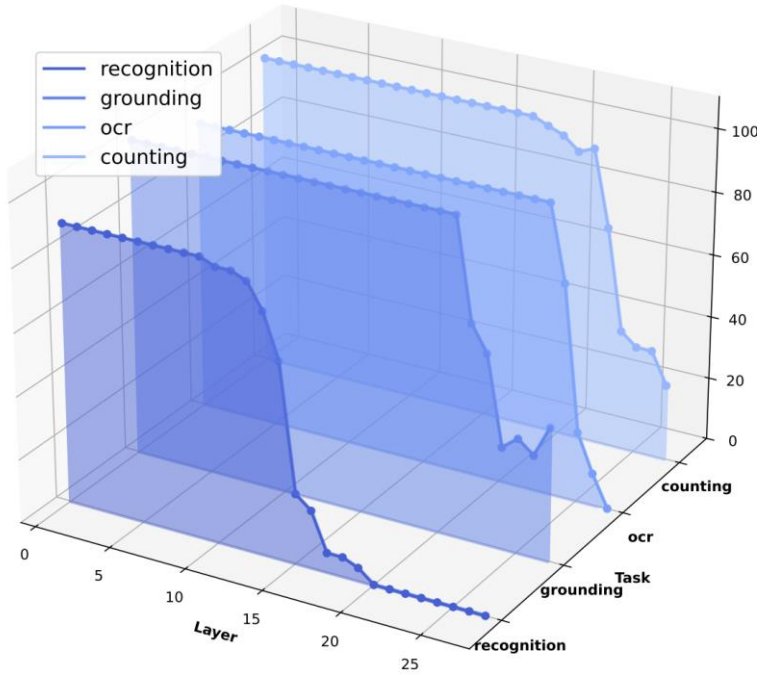


$$\mathcal{A}^{(\ell)}(\mathcal{R}\to\mathcal{V}) = \frac{1}{|\mathcal{H}||\mathcal{R}||\mathcal{V}|}\sum_{h\in\mathcal{H}}\sum_{i\in\mathcal{R}}\sum_{j\in\mathcal{V}} A_{ij}^{(\ell,h)}.$$

Layer-wise functional localization of vision

Construct paired images that differ in exactly one visual attribute





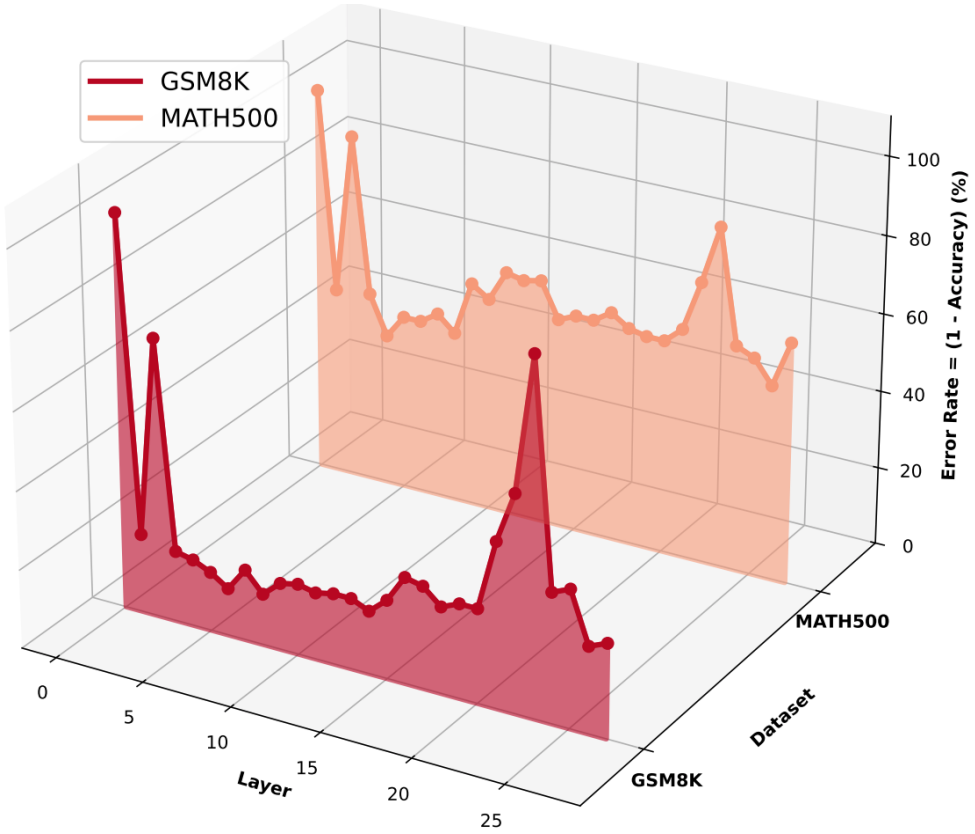
(a) Vision localization via vision-token swapping.

$$\text{Change Rate}^{(\ell)} = \frac{1}{N}\sum_{n=1}^{N}\mathbb{I}\left[f(i_n^{(\ell)}, p_n) \neq f(i_n'^{(\ell)}, p_n)\right],$$

Findings: Vision-related functionalities are primarily associated with Early and Mid transformer layers.

Layer-wise functional localization of reasoning

Directly removes the contribution of a specific transformer layer.



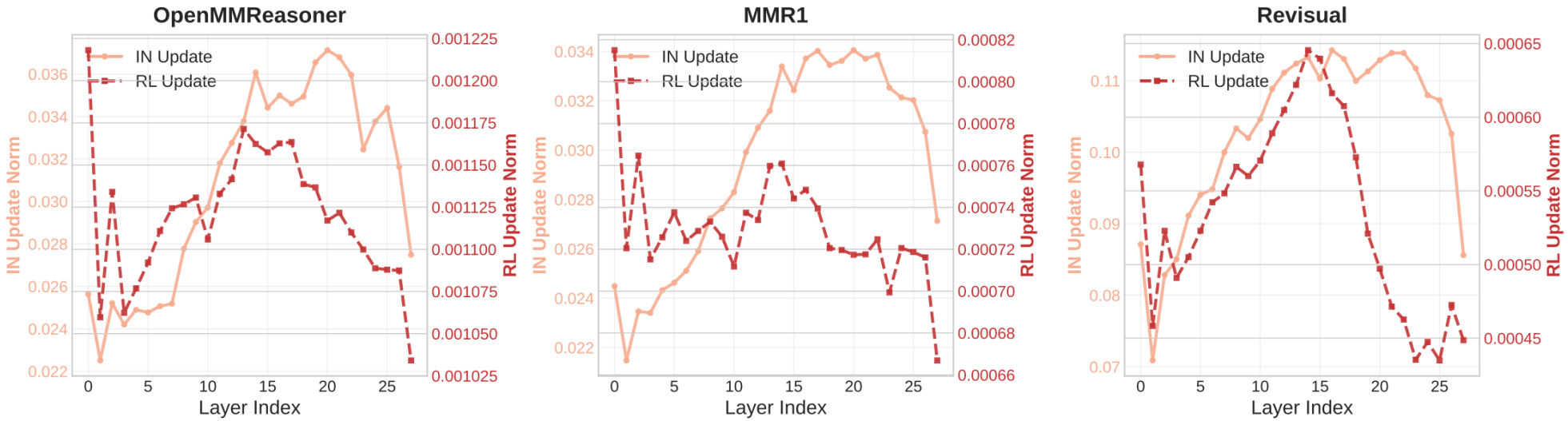
(b) Reasoning localization via layer skipping.

Finding: reasoning-related computations are concentrated in Late layers.

Update Energy

Characterize the post-training changes through the geometry of parameter updates.

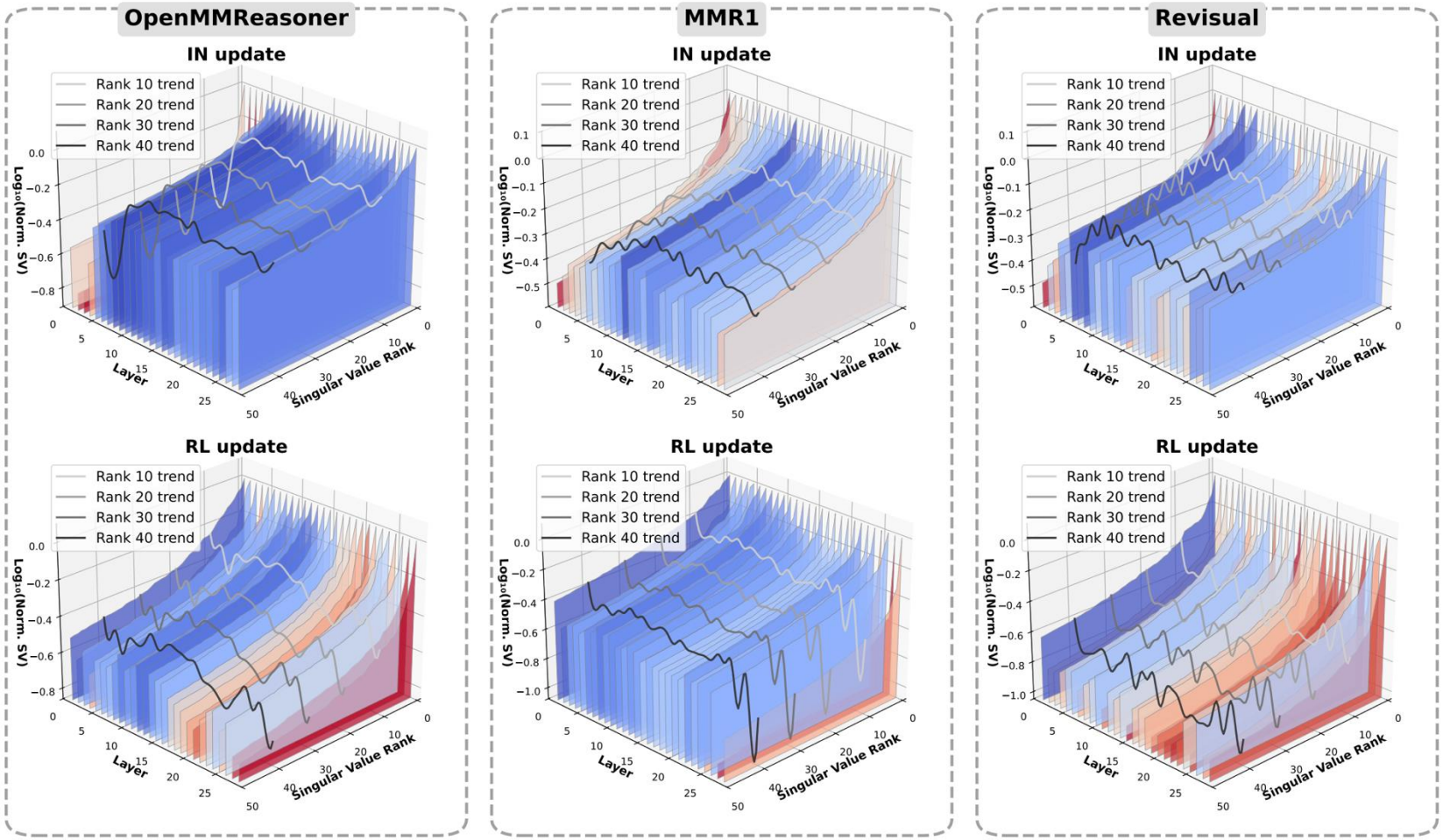
$$\Delta \mathbf{W}^{(\ell)} = \mathbf{W}^{(\ell)}_{\text{trained}} - \mathbf{W}^{(\ell)}_{\text{base}} \qquad \|\Delta \mathbf{W}^{(\ell)}\|_F$$



Finding; Both IN and RL impose high norm update in Mid layers.

Update Diversity

$$\Delta \mathbf{W}^{(\ell)} = \mathbf{U}\,\mathrm{diag}(\sigma_1, \ldots, \sigma_r)\,\mathbf{V}^\top, \qquad \log(\sigma_i/\sigma_1).$$



IN and RL impose complementary optimization diversity: RL applies less diverse refinements to Mid-Late layers.

Transferability Test via Model Merging

Ability	Base	IN	Region-wise Merged Models						RL
			IN : RL : RL	IN : IN : RL	RL : RL : IN	RL : IN : IN	IN : RL : IN	RL : IN : RL	
Training Recipe – OpenMMReasoner (Zhang et al., 2025c)									
Vision	38.0	47.0	42.0 *(-5.0)*	48.0 *(+1.0)*	45.0 *(-2.0)*	45.0 *(-2.0)*	44.0 *(-3.0)*	40.0 *(-7.0)*	42.0
Vision-to-Reasoning	46.0	55.0	59.0 *(+4.0)*	58.0 *(+3.0)*	63.0 *(+8.0)*	57.0 *(+2.0)*	58.0 *(+4.0)*	55.0 *(-0.0)*	61.0
Reasoning	63.0	78.0	81.0 *(+3.0)*	73.0 *(-5.0)*	78.0 *(-0.0)*	73.0 *(-5.0)*	80.0 *(+2.0)*	73.0 *(-5.0)*	78.0
Training Recipe – MMR1 Leng et al. (2025)									
Vision	38.0	44.0	37.0 *(-7.0)*	41.0 *(-3.0)*	45.0 *(+1.0)*	45.0 *(+1.0)*	41.0 *(-3.0)*	43.0 *(-1.0)*	50.0
V-to-R	46.0	46.0	51.0 *(+5.0)*	50.0 *(+4.0)*	50.0 *(+4.0)*	56.0 *(+10.0)*	52.0 *(+6.0)*	57.0 *(+13.0)*	54.0
Reasoning	63.0	60.0	61.0 *(+1.0)*	64.0 *(+4.0)*	64.0 *(+4.0)*	79.0 *(+19.0)*	56.0 *(-4.0)*	72.0 *(+12.0)*	62.0
Training Recipe – Revisual Chen et al. (2025c)									
Vision	38.0	40.0	42.0 *(+2.0)*	42.0 *(+2.0)*	41.0 *(+1.0)*	37.0 *(-3.0)*	40.0 *(-0.0)*	46.0 *(+6.0)*	35.0
Vision-to-Reasoning	46.0	56.0	59.0 *(+3.0)*	58.0 *(+2.0)*	58.0 *(+2.0)*	56.0 *(-0.0)*	60.0 *(+4.0)*	51.0 *(-5.0)*	59.0
Reasoning	63.0	81.0	85.0 *(+4.0)*	84.0 *(+3.0)*	84.0 *(+3.0)*	85.0 *(+4.0)*	86.0 *(+5.0)*	88.0 *(+7.0)*	88.0

Model merging confirms consistent RL refinements in Mid-Late layers, which consistently improve vision-to-reasoning alignment and reasoning capabilities.

Validation via Model Freezing

	Fine-grained Metrics			Benchmark Metrics		
	Vision (M_{vis})	**Vision-to-Reasoning (M_{v2r})**	**Reasoning (M_{rea})**	**MathVista**	**MathVision**	**MathVerse**
	Original Training Recipe					
IN Model (Zhang et al., 2025c)	34.0	21.0	26.0	46.5	18.4	37.0
RL Model (Zhang et al., 2025c)	33.0	29.0	34.0	48.1	14.1	37.8
	Training with Frozen Blocks					
RL Model - Frozen Early Block	**35.0**	**31.0**	36.0	**48.2**	**21.0**	34.5
RL Model - Frozen Mid Block	25.0	29.0	**38.0**	46.5	15.5	**35.7**
RL Model - Frozen Late Block	30.0	27.0	34.0	47.9	16.8	35.0

RL gains depend causally on Mid-Late refinement: Preventing RL from updating the Late region during training eliminates improvements in both fine-grained and benchmark metrics.

TRUST-REGION ADAPTIVE POLICY OPTIMIZATION

Mingyu Su[1], **Jian Guan**[2], **Yuxian Gu**[1], **Minlie Huang**[1], **Hongning Wang**[1*]
[1]The Conversational AI (CoAI) Group, Tsinghua University [2]Ant Group
sumy25@mails.tsinghua.edu.cn, hw-ai@tsinghua.edu.cn

Setting

Better RL method for LLM reasoning

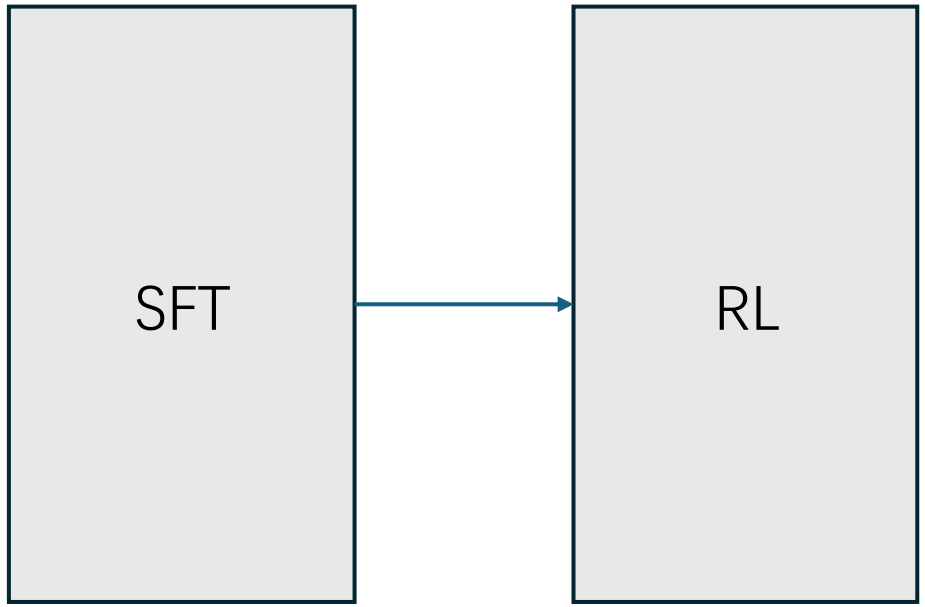
Question

The polynomial $x^3 - 3x^2 + 4x - 1$ is a factor of $x^9 + px^6 + qx^3 + r$. Enter the ordered triple (p, q, r).
Answer: (6,31,-1)

TRAPO

[Clear Plan and Sufficient and moderate contemplation]
Sure, I can help you solve this problem. The goal is to find the ordered triple \((p, q, r)\) such that the polynomial \(x^3 - 3x^2 + 4x - 1\) is a factor of \(x^9 + px^6 + qx^3 + r\).
Step 1: Understand the Problem
Given that \(x^3 - 3x^2 + 4x - 1\) is a factor of \(x^9 + px^6 + qx^3 + r\), this implies that when we divide …
Step 2: Find the Roots of the Cubic Polynomial
First, let's find the roots of the cubic polynomial \(x^3 - 3x^2 + 4x - 1 = 0\). This might be a bit tricky,…
Step 3: Polynomial Division or Substitution
Since \(x^3 - 3x^2 + 4x - 1\) is a factor of \(x^9 + px^6 + qx^3 + r\), we can use polynomial division or substitute-ion. However, polynomial division might be cumbersome for degree 9. Instead, let's use substitution…
Step 4: Substitute and Simplify
Let's compute \(y^3 + py^2 + qy + r\) with \(y = 3x^2 - 4x + 1\)…
Step 5: Express Higher Powers in Terms of Lower Powers
We need to express \(x^9\) in terms of \(x^2\), \(x\), and constants using the relation \(x^3 = 3x^2 - 4x + 1\)…
Step 6: Substitute into the Original Polynomial
Now, substitute these expressions into the original polynomial \(x^9 + px^6 + qx^3 + r\):…
Step 7: Solve the System of Equations
First, let's solve the first two equations for \(p\) and \(q\)…
Step 8: Verify the Solution
To ensure that our solution is correct, we can verify by substituting \(p = 6\), \(q = 31\), \(r = -1\) back into the original polynomial and checking if \(x^3 - 3x^2 + 4x - 1\) divides it.
Conclusion
The ordered triple \((p, q, r)\) is \((6, 31, -1)\). Therefore, the answer is \(\boxed{(6, 31, -1)}\).
Reflection
This problem required understanding polynomial division, substitution, and solving a system of equations. The key was to express higher powers of \(x\) in terms of lower powers using the given cubic …

Motivation

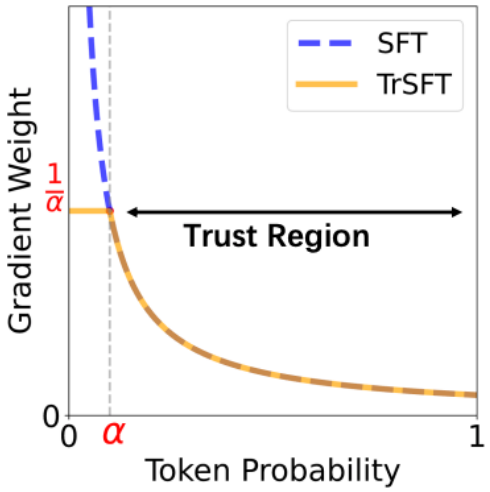


The dominant two-stage pipeline (SFT → RL) is internally inconsistent: SFT enforces rigid imitation that can **suppress exploration** and **induce forgetting**, limiting the gains achievable by RL.

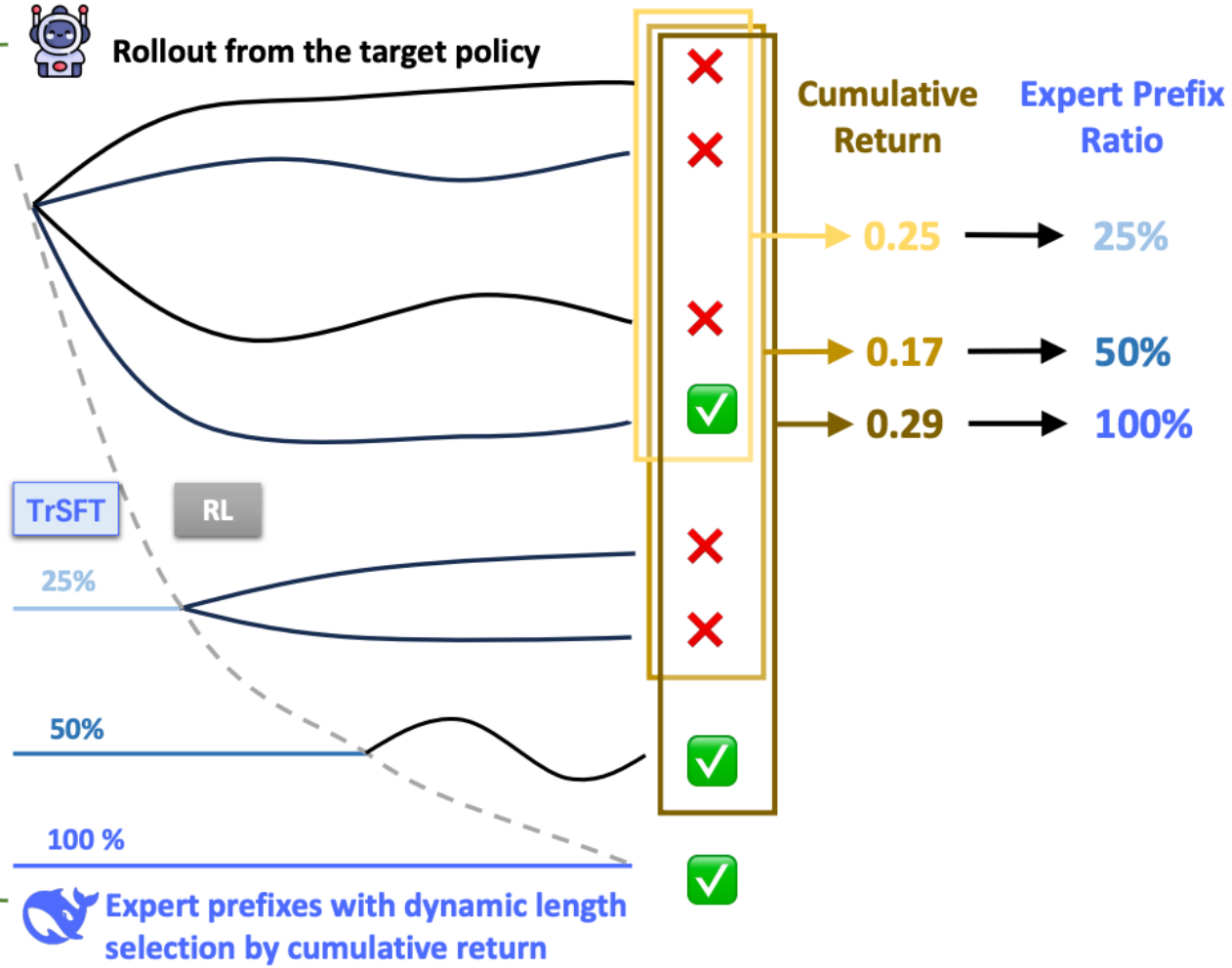
Method

A overview …

Objective	Gradient
SFT	$\frac{1}{p^\theta(y_n \mid \boldsymbol{x}, \boldsymbol{y}_{<n})} \nabla_\theta p^\theta(y_n \mid \boldsymbol{x}, \boldsymbol{y}_{<n})$
TrSFT	$\frac{1}{\max(p^\theta(y_n \mid \boldsymbol{x}, \boldsymbol{y}_{<n}), \alpha)} \nabla_\theta p^\theta(y_n \mid \boldsymbol{x}, \boldsymbol{y}_{<n})$



Find the volume of space obtained by rotating the region bounded by $y = \ln x$, y=0, $x \in [0, e^2]$

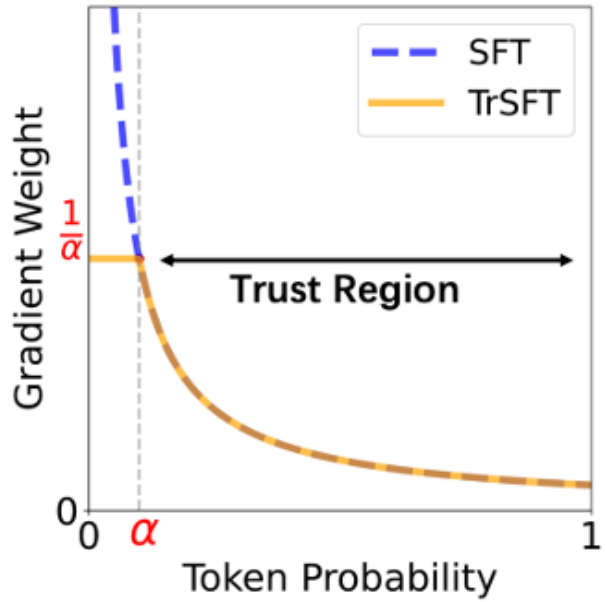


Expert Model Target Policy Model

Trust-Region SFT

Problem: Standard SFT can create distribution blending, pushing probability mass into "void regions" unsupported by either expert or current policy

Fix: Clip the SFT gradient weight with a trust-region parameter α, prevent exploding gradients on low-probability tokens.



(alpha=0.1).

$$L_{\mathrm{SFT}}(\theta) = \mathbb{E}_{\boldsymbol{x}\sim X}\left[\mathbb{E}_{\boldsymbol{y}\sim p_{\mathrm{E}}(\cdot|\boldsymbol{x})}\left[-\log p_{\mathrm{T}}^{\theta}(\boldsymbol{y}|\boldsymbol{x})\right]\right].$$

$$\nabla_\theta L_{\mathrm{SFT}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{n=1}^{|\boldsymbol{y}^i|}\frac{1}{p_{\mathrm{T}}^{\theta}(y_n^i|\boldsymbol{x}^i,\boldsymbol{y}^i_{<n})}\nabla_\theta p_{\mathrm{T}}^{\theta}(y_n^i|\boldsymbol{x}^i,\boldsymbol{y}^i_{<n}).$$

$$\nabla_\theta L^{\alpha}_{\mathrm{TrSFT}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{n=1}^{|\boldsymbol{y}^i|}\frac{1}{\max\left(p_{\mathrm{T}}^{\theta}(y_n^i|\boldsymbol{x}^i,\boldsymbol{y}^i_{<n}),\alpha\right)}\nabla_\theta p_{\mathrm{T}}^{\theta}(y_n^i|\boldsymbol{x}^i,\boldsymbol{y}^i_{<n}),$$

Micro-group sampling

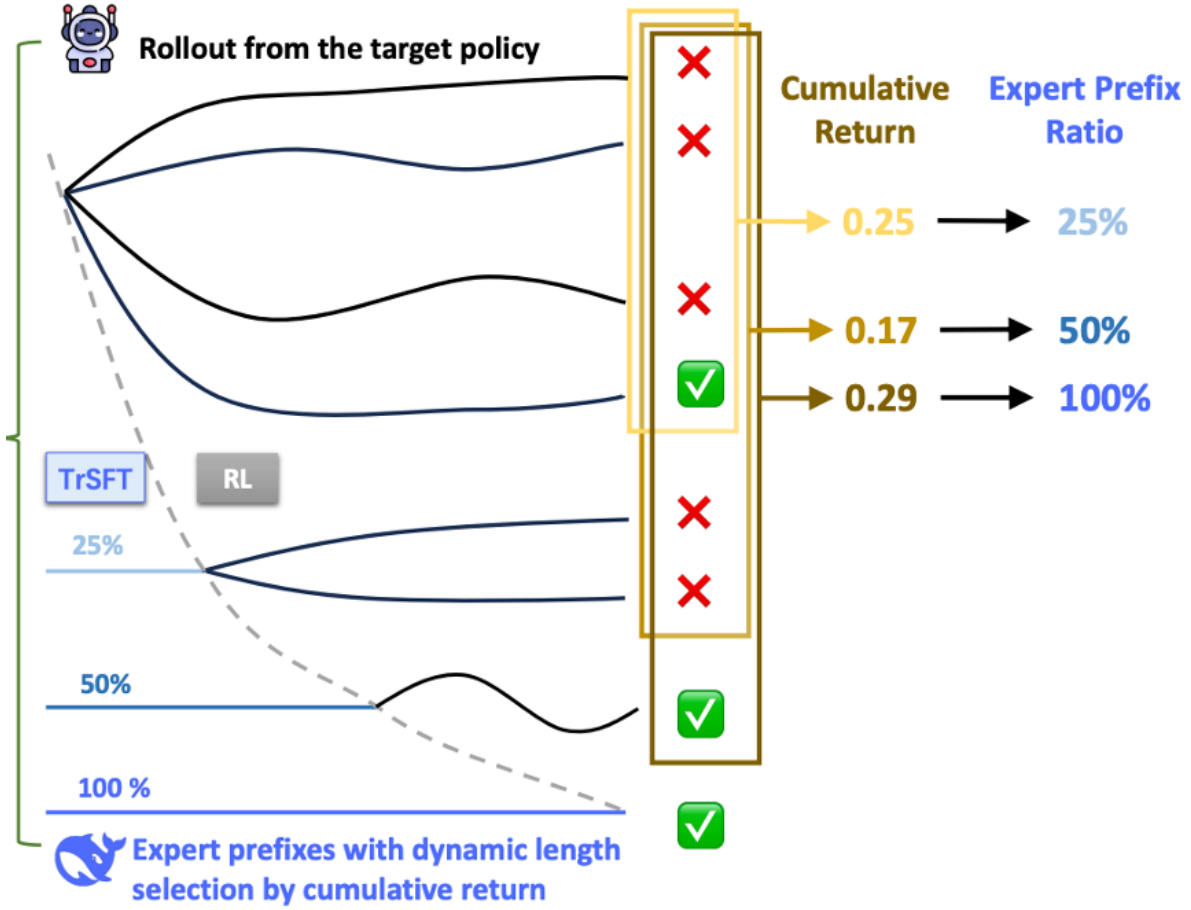
TRAPO adaptively allocates guidance using **ordered micro-groups:**

1. **Split rollouts** into groups with a prefix ratio.

2. Start with **no guidance**;

3. if the average return is below the threshold, **provide a longer expert prefix in later micro-groups**; otherwise, keep exploring without prefixes.

Config:
Group size 8 split into {4,2,1,1}, with (L=(0,0.2,0.5,1.0)), thresholds (-1,0.5,0.7,0.9).



Result

Model	Mathematical Reasoning						General Domain Reasoning		
	AIME2024	AMC	MATH-500	Minerva	Olympiad	Avg.	ARC-c	MMLU-Pro	Avg.
Qwen2.5-Math-7B	11.9	33.7	47.0	12.5	21.9	25.4	34.5	23.3	28.9
Qwen2.5-Math-7B-Instruct	11.3	48.2	82.6	36.8	39.7	43.7	72.4	38.3	55.4
Pure RL without External Expert Guidance									
GRPO	24.0	59.0	84.0	39.3	45.8	50.4	80.5	47.2	63.9
PRIME-Zero*	17.0	54.0	81.4	39.0	40.3	46.3	73.3	32.7	53.0
SimpleRL-Zero*	27.0	54.9	76.0	25.0	34.7	43.5	30.2	34.5	32.4
OpenReasoner-Zero*	16.5	52.1	82.4	33.1	47.1	46.2	66.2	**58.7**	62.5
Oat-Zero*	33.4	61.2	78.0	34.6	43.4	50.1	70.1	41.7	55.9
RL Incorporating External Expert Guidance									
SFT	27.7	56.0	84.8	38.2	44.7	50.3	51.5	33.1	42.3
SFT-then-RL	**33.5**	62.3	86.6	41.2	47.6	54.3	52.9	36.1	44.5
ReLIFT	28.2	64.9	87.4	33.8	52.5	53.4	76.2	52.5	64.4
LUFFY	29.4	65.5	88.4	38.2	56.0	55.5	80.5	53.0	66.7
Our Method									
TRAPO	28.3	**66.2**	**89.2**	**41.5**	**57.6**	**56.6**	**83.7**	52.8	**68.3**

Ablation

Table 2: Ablation study on TRAPO components.

Model	AIME2024	AMC	MATH-500	Minerva	Olympiad	Avg.
GRPO (Qwen2.5-Math-7B)	24.0	59.0	84.0	39.3	45.8	50.4
+ Micro-group sampling	26.0	63.7	84.2	39.3	50.1	52.7
+ Micro-group sampling + *SFT Loss*	14.6	32.8	60.2	25.4	28.5	32.3
+ Micro-group sampling + *LUFFY Loss*	26.2	65.1	87.2	36.8	52.5	53.6
+ Micro-group sampling + *TrSFT Loss*	28.3	66.2	89.2	41.5	57.6	56.6