Dive into RL

Bae Sun Woo

Feb 9, 2026

Table of Contents

ProRL

“Does reinforcement learning truly unlock new reasoning capabilities from a base model, or does it merely optimize the sampling efficiency of solutions already embedded in the base model?”

ProRL: Nemotron-Research-Reasoning-Qwen-1.5B

Deepseek R1-1.5B	+ GRPO(+DAPO techniques)	+ **KL Divergence**	+ **Periodic Reference Model Reset**
Base	Clip Higher & Dynamic Sampling	Stability	“Key” of Prolonged Learning

Nemotron-Research-Reasoning-Qwen-1.5B

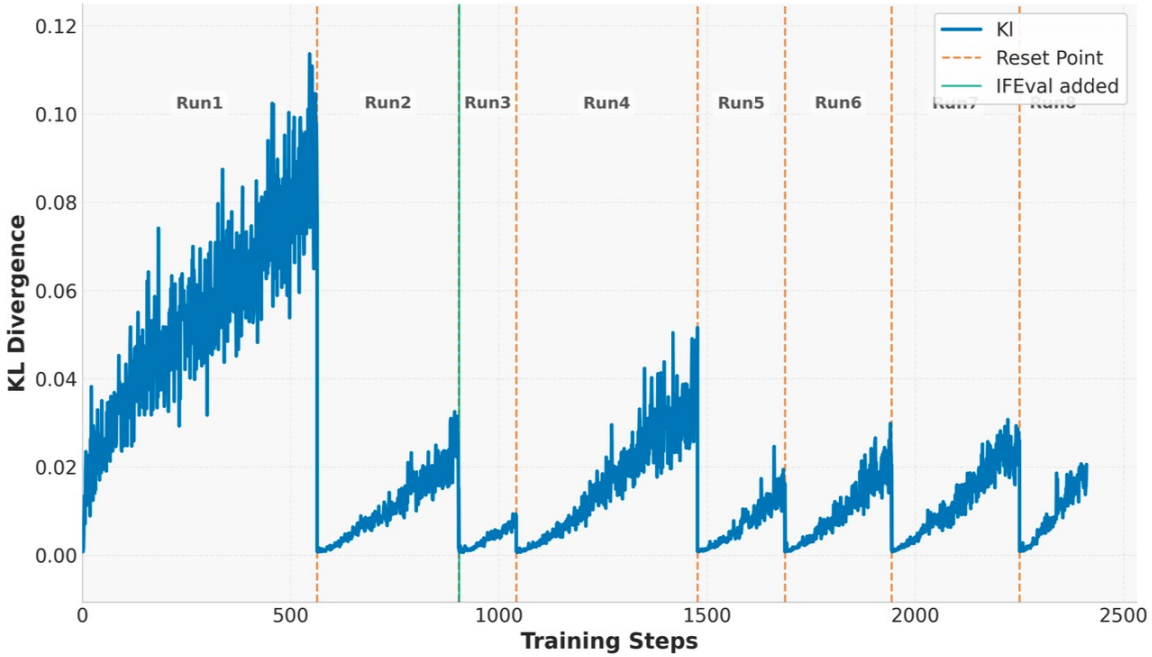
ProRL: KL Divergence & Periodic Reference Model Reset

$$L_{KL-RL}(\theta) = L_{GRPO}(\theta) - \beta D_{KL}(\pi_\theta || \pi_{ref})$$

Recent works have excluded KL Divergence
=> base model prior SFT

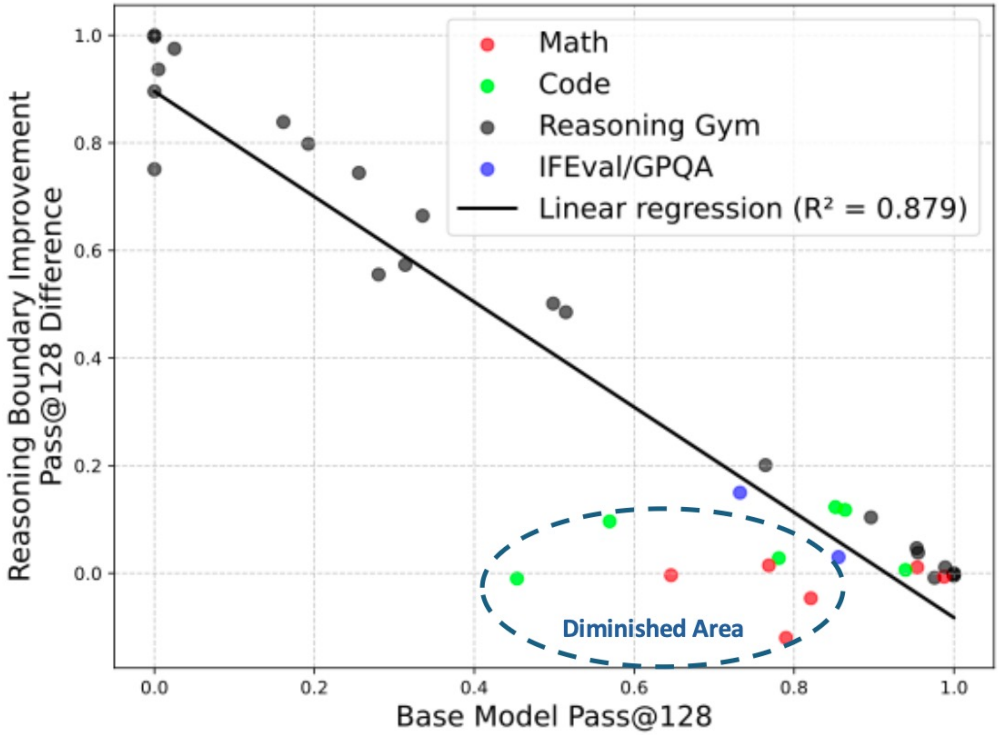
Here, base mode is well-initialized checkpoint (DeepSeek-R1-Distill-Qwen-1.5B)
=> Bring KL divergence for both stability and sustained entropy

However, KL term may increasingly dominate the loss, leading to diminishing policy updates



=> Bring Reference Model Reset => Continue to improve

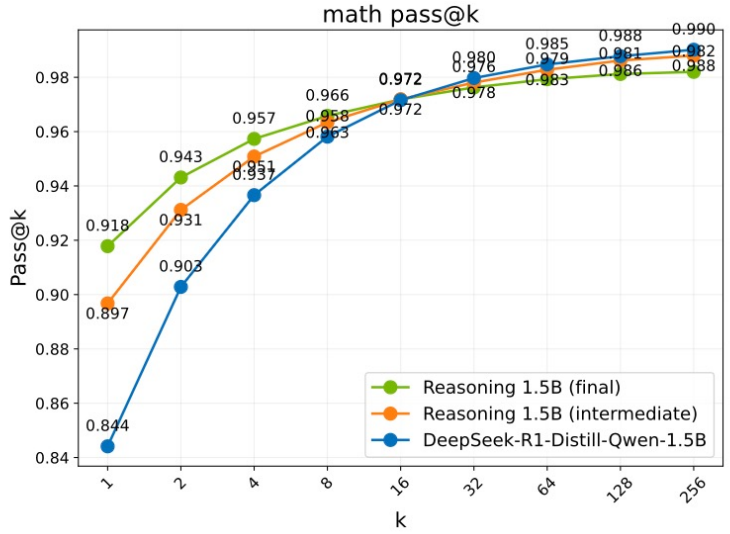
Insight 1:The Weaker the Start, the Stronger the Gain with ProRL

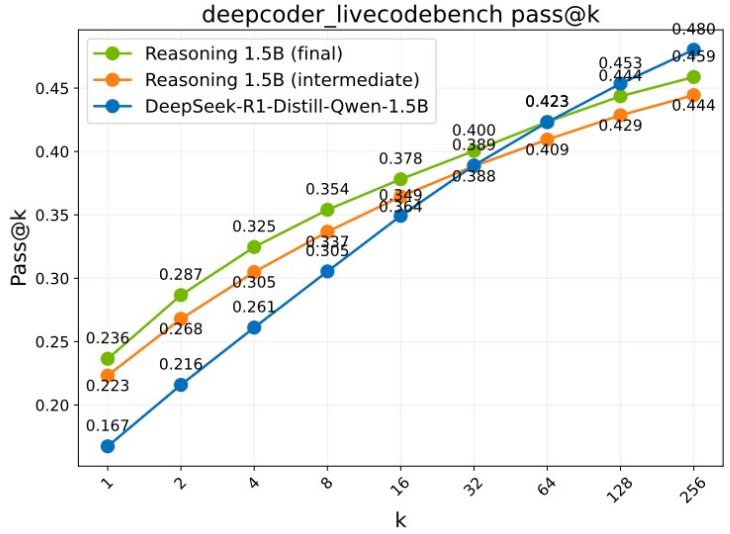


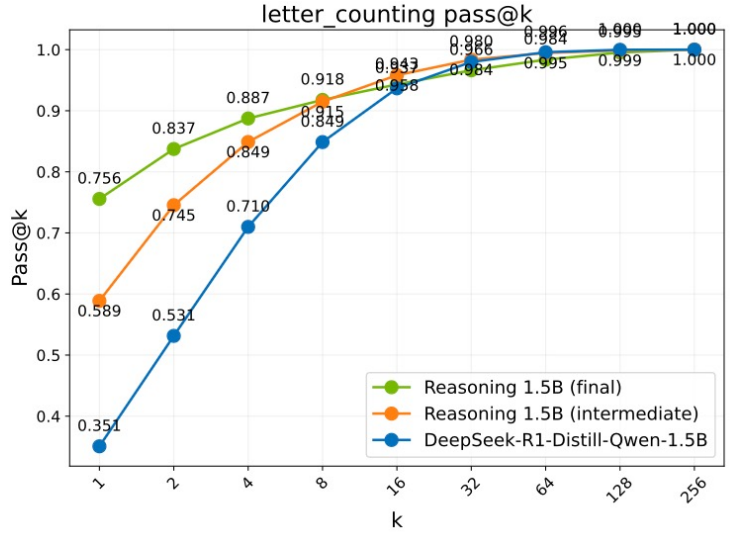
Negative correlation between the base model's reasoning boundary and the extent of reasoning improvement after RL training

Insight 2-(1): ProRL's Reasoning Boundaries: **Diminish**, Plateau, and Sustained Gains

Diminish





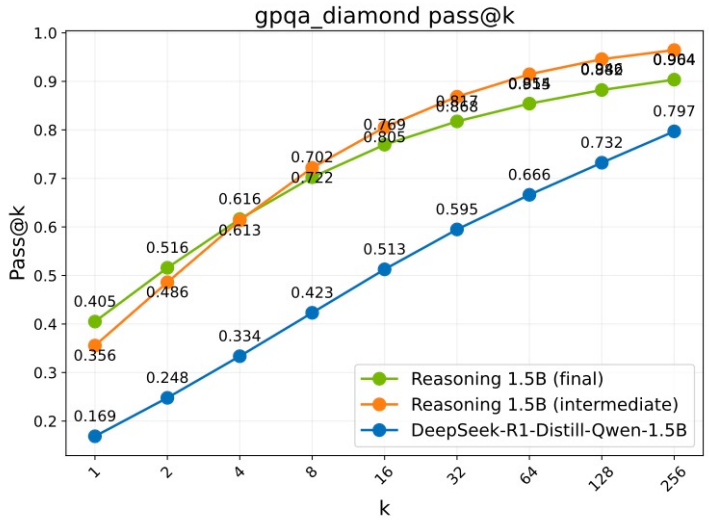


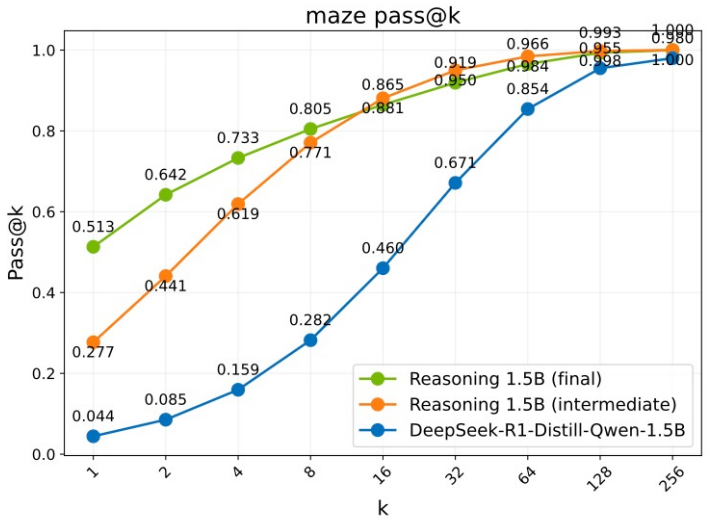
pass@128 : reflect broader reasoning ability

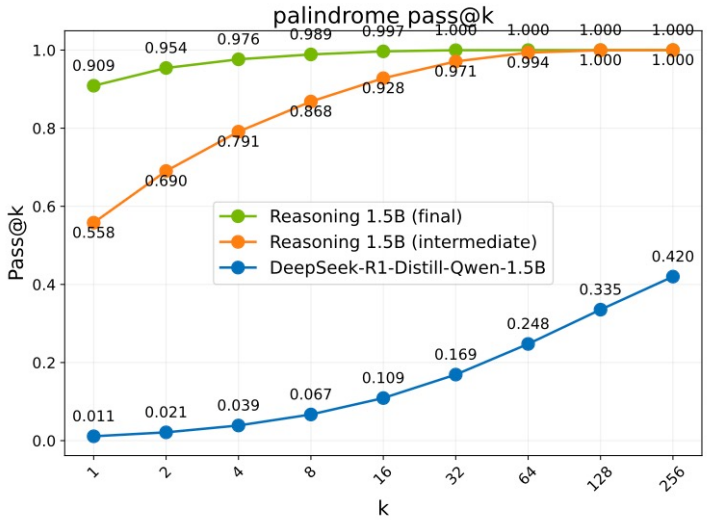
"Decreased or unchanged reasoning capacity"

Insight 2-(2): ProRL's Reasoning Boundaries: Diminish, **Plateau**, and Sustained Gains

Plateau





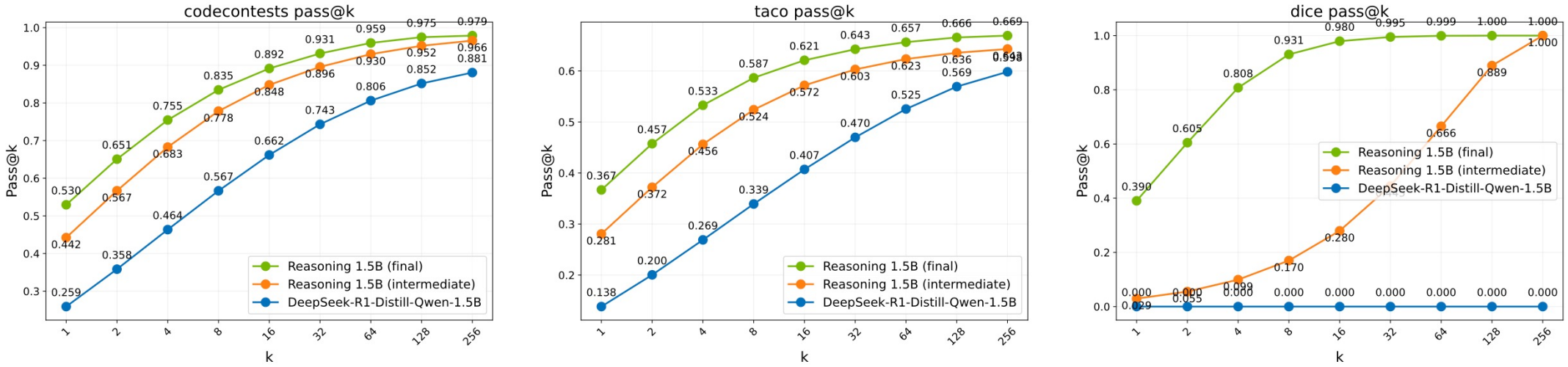


Increase in both pass@1 and pass@128

However, **these gains are largely achieved early in training**

Insight 2-(3): ProRL's Reasoning Boundaries: Diminish, Plateau, and **Sustained Gains**

Sustained



Continued improvements in reasoning capacity

Tasks which require extensive exploration

Insight 3: Enhances Out-of-Distribution Reasoning

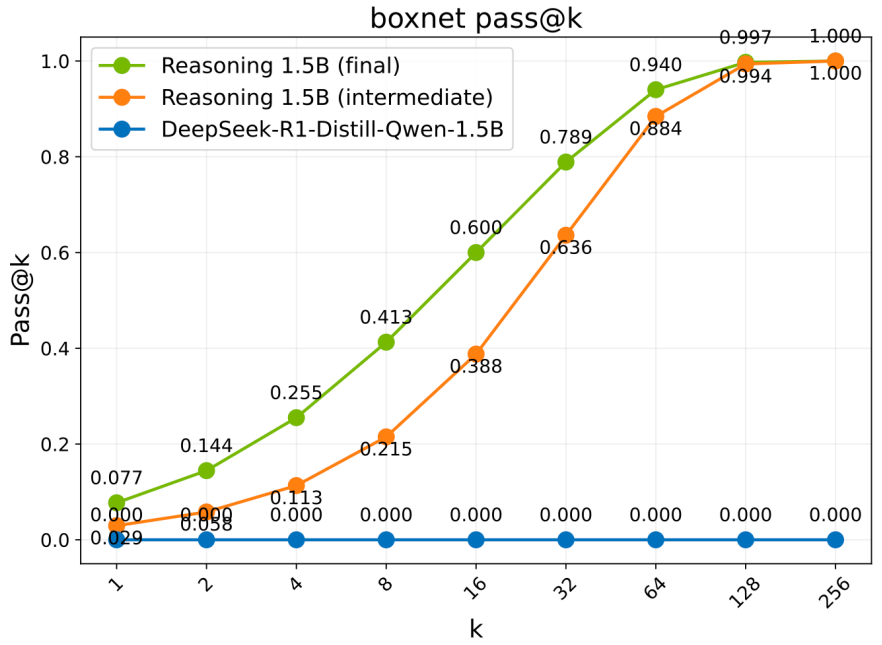
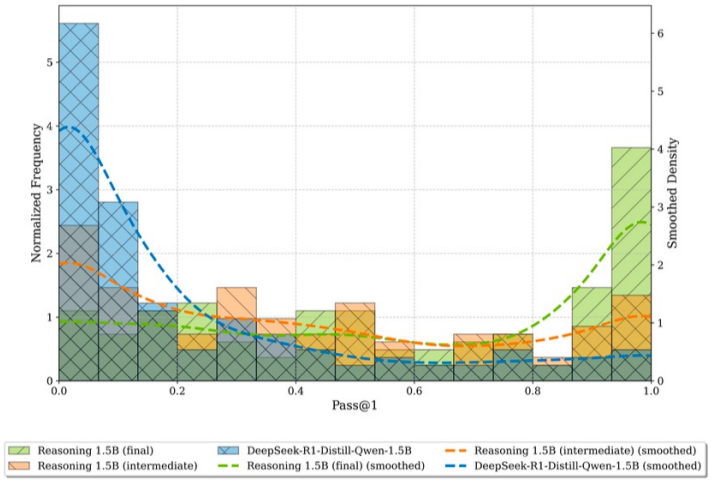


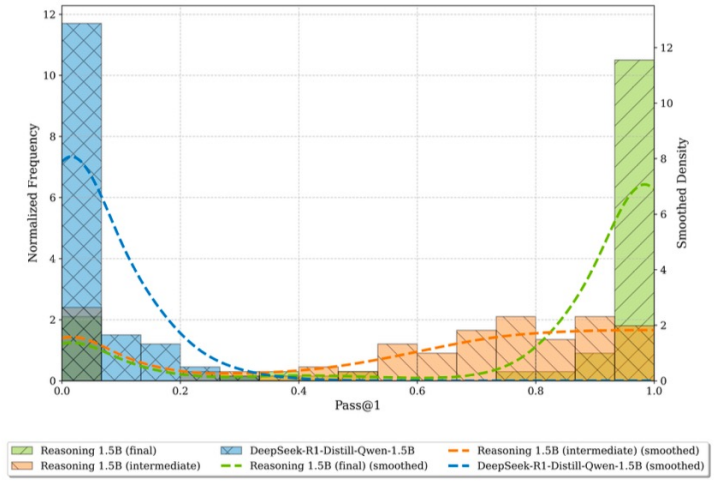
Figure 5: Expanded reasoning boundary for OOD task *boxnet*.

Not seen during training

Insight 4: pass@1 Distributions Evolution



(a) Codeforces



(b) *family_relationships*

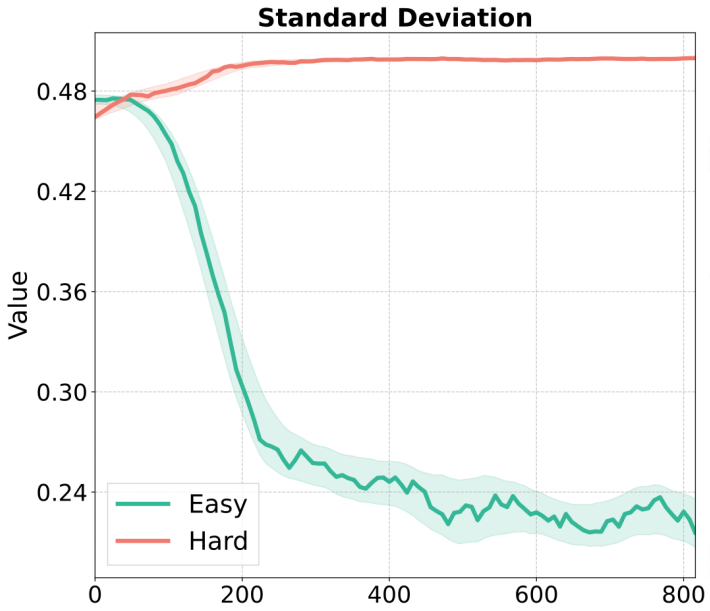
Figure 7: Distribution shifts in pass@1 accuracy following prolonged RL training across two representative tasks. The figure illustrates the evolution of pass@1 probability distributions for selected tasks from code (a) codeforces, and reasoning domains (b) *family_relationships*.

Part I: Tricks or Traps? A Deep Dive into RL for LLM Reasoning

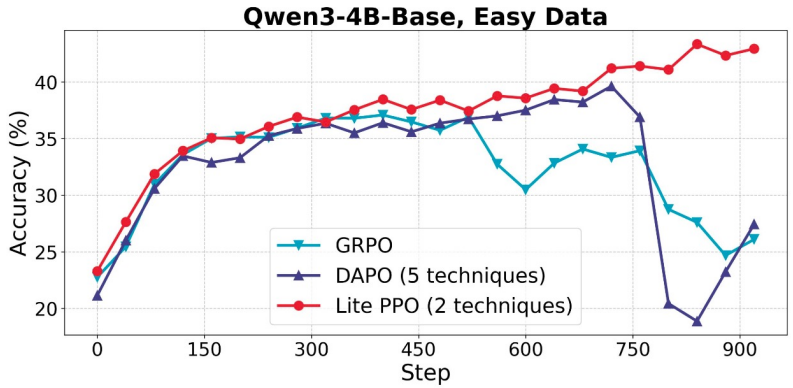
"Absence of standardized guidelines for applying RL techniques and a fragmented understanding of their underlying mechanisms"

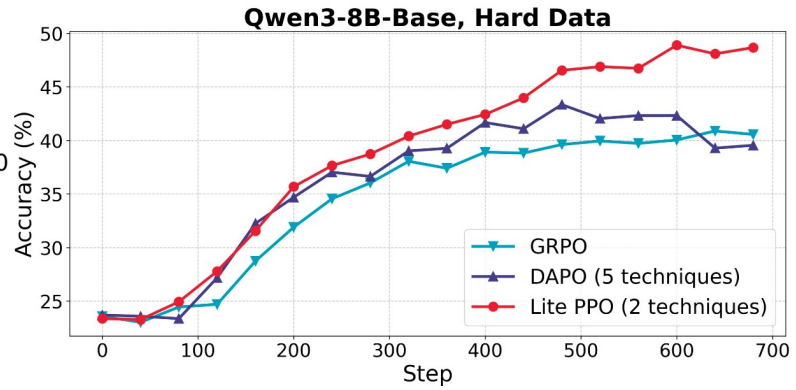
1. Normalization
2. Clip-Higher
3. Loss Aggregation
4. Overlong Filtering

Analysis 1: Normalization



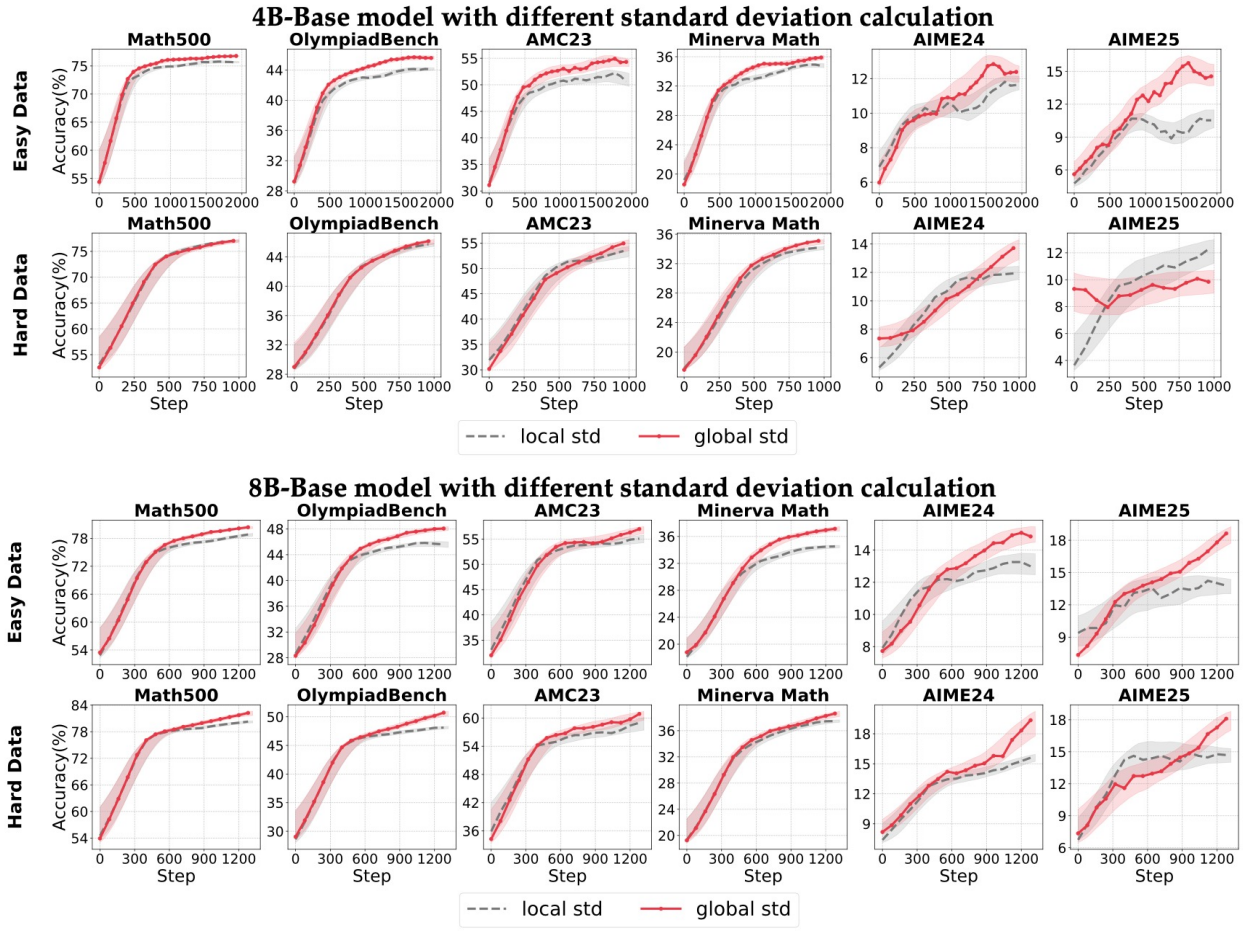
$$A_k^{\text{group}} = \frac{r_k - \text{mean}(\{r_j\}_{j=1}^{K})}{\text{std}(\{r_j\}_{j=1}^{K})}$$



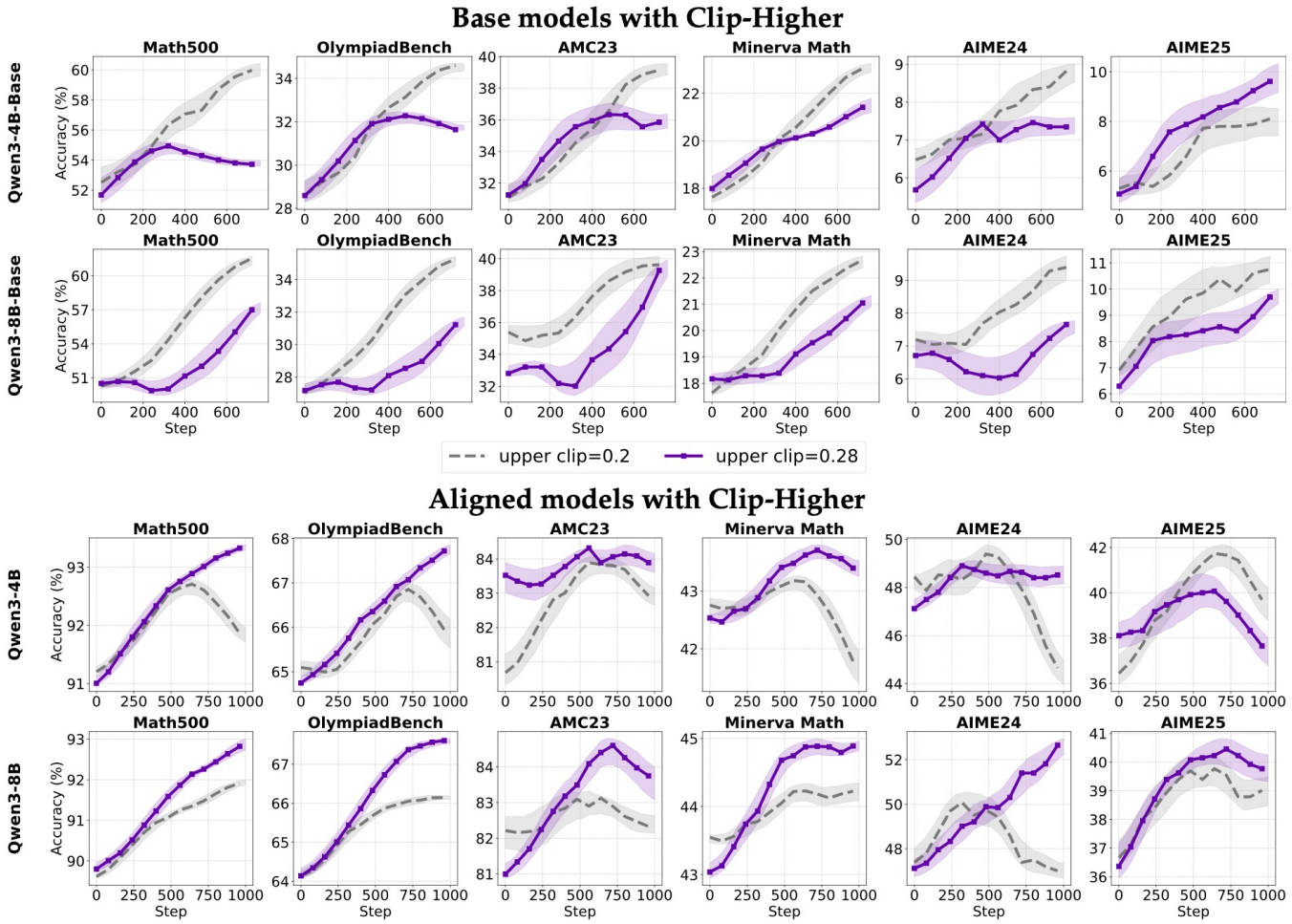


Mean: local(group) level
Std: global(batch) level)

Analysis 1: Normalization



Analysis 2: Clip-Higher – Base vs Aligned



CH(Base) < CH(Aligned)

Why?

Clipping rate of Base model: 0.003

Naive policy expressiveness hinders exploration
=> No use

Analysis 2: Clip-Higher – Small vs Large

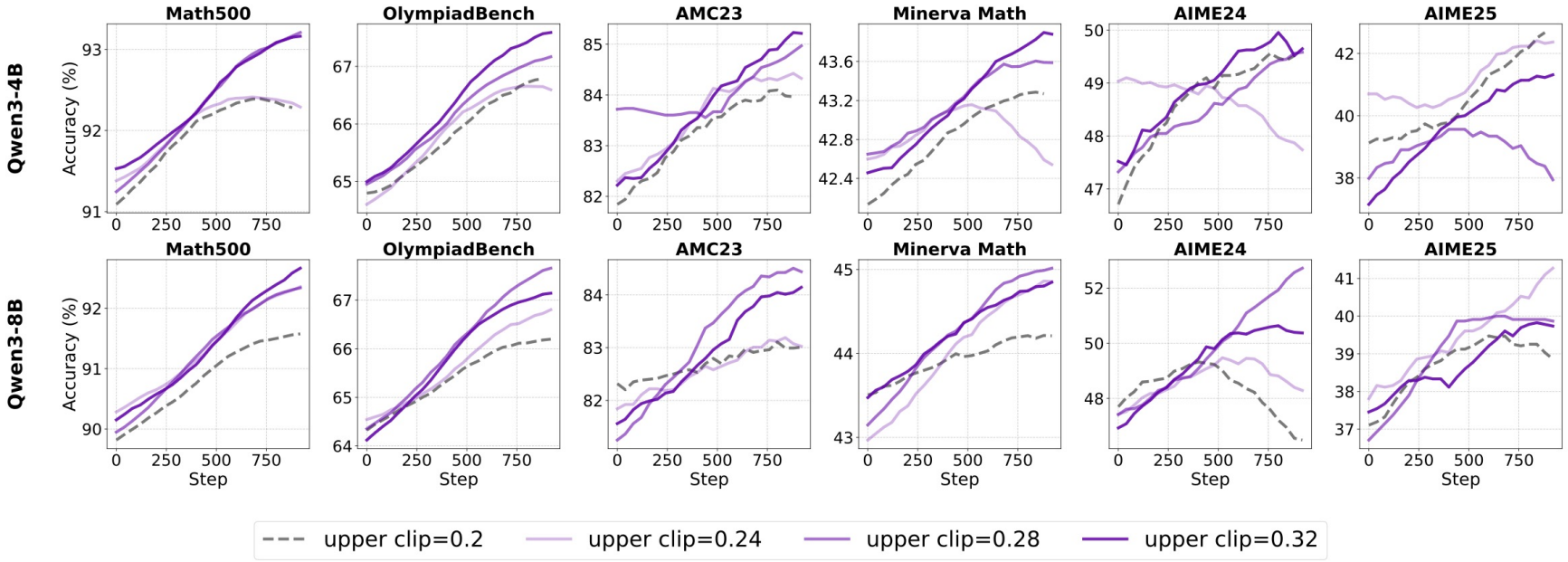


Figure 11: Test accuracy of aligned models (trained on medium data) with various clipping upper bounds.

Small model: Accuracy(0.32) > Accuracy(0.28)
Larger model: Accuracy(0.32) < Accuracy(0.28)

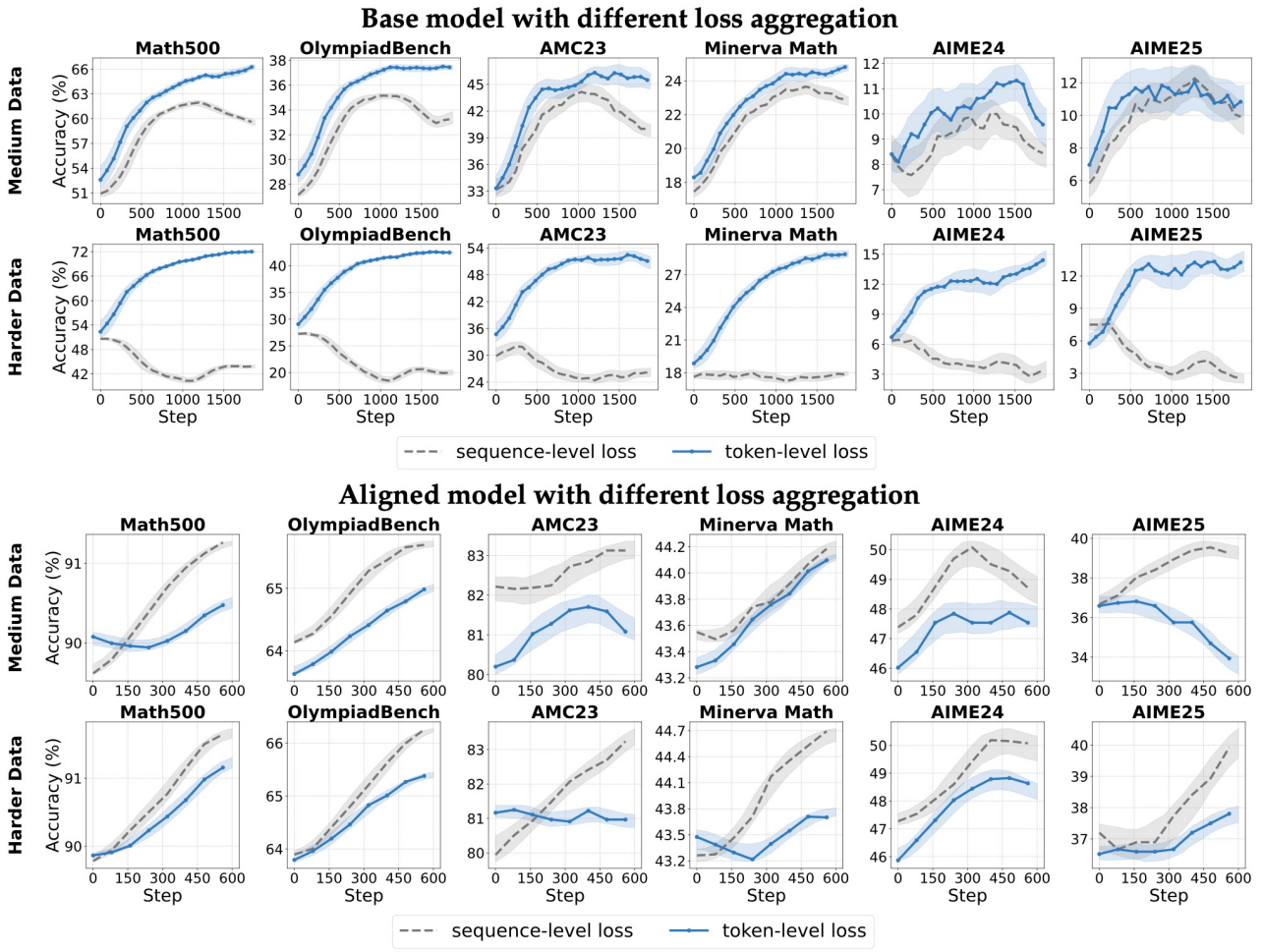
Analysis 3: Loss Aggregation – seq vs token

$$\mathcal{J}_{\text{sequence-level}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\left(r_{i,t}(\theta),\ 1-\epsilon_{\text{low}},\ 1+\epsilon_{\text{high}}\right)\hat{A}_{i,t}\right)\right]$$

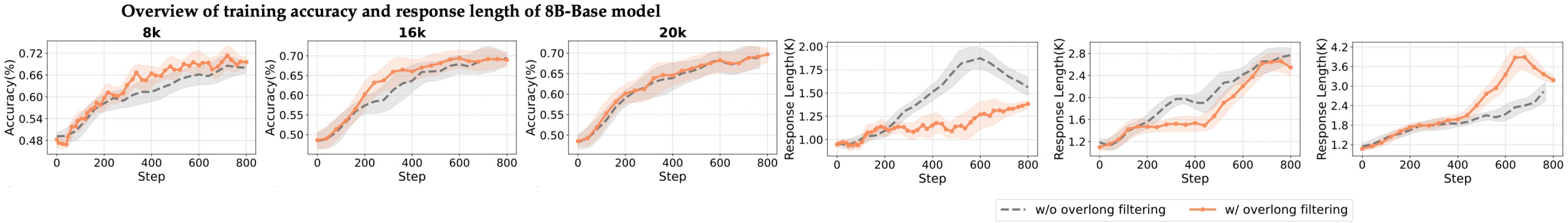
$$\mathcal{J}_{\text{token-level}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\left(r_{i,t}(\theta),\ 1-\epsilon_{\text{low}},\ 1+\epsilon_{\text{high}}\right)\hat{A}_{i,t}\right)\right]$$

Base model: Token-level loss

Aligned model: Sequence-level loss



Analysis 4: Overlong Filtering – long vs short



	Short Threshold (8k)	Threshold 20k
What primarily filtered	Samples that are long due to extended reasoning	Unproductive or "negative" samples that contribute little to model learning
Model will	produce shorter, more concise responses, discouraging excessive verbosity	generate longer responses in comparison to the vanilla policy.
Benefits	substantial	diminished

“Meanwhile, if practitioners expect that LLMs generate extremely long reasoning paths, data mask may remove pathological outputs without significantly affecting valid reasoning sequences”

Simple Combination: Lite PPO

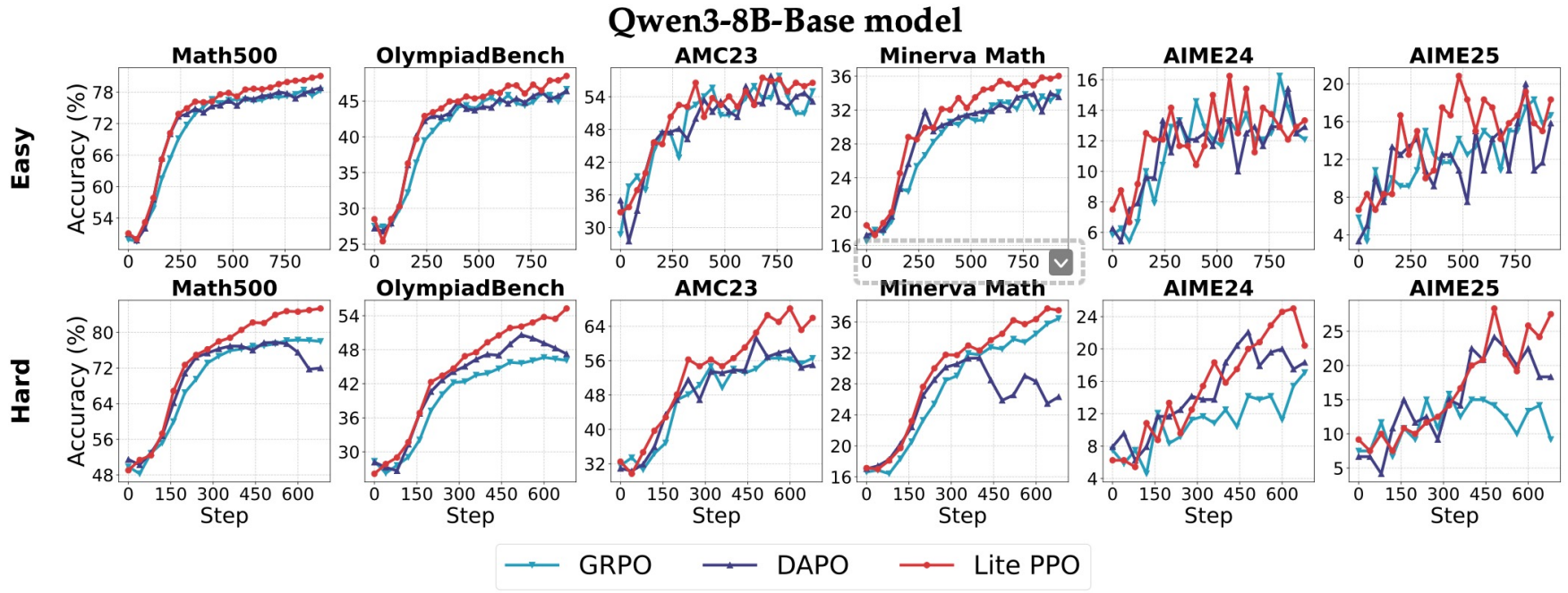


Figure 15: Test accuracy of non-aligned models trained with three RL methods, i.e., Lite PPO (ours), GRPO (Shao et al., 2024) and DAPO (Yu et al., 2025).

Normalization: *group-level mean* calculation and *batch-level standard deviation* calculation

+

Token-level loss aggregation emerges as another highly effective technique for non-aligned models

JustRL: Scaling a 1.5B LLM with a Simple RL Recipe

“Multi-stage training pipelines, dynamic hyperparameter schedules, adaptive temperature controls…Is this complexity necessary?

JustRL: RL Comparisons

Model	EC	THP	TTP	RKL	LC	AT	RR	DS	ST	Date
STILL-3-1.5B	✗	✓	✓	✓	✗	✗	✗	✗	✗	Jan '25
DeepScaleR-1.5B	✓	✗	✗	✗	✓	✗	✗	✗	✓	Feb '25
FastCuRL-1.5B	✗	✓	✗	✗	✓	✗	✗	✗	✓	Mar '25
ProRL-V1	✓	✓	✗	✓	✓	✗	✗	✓	✓	May '25
e3-1.7B	✓	✓	✗	✗	✓	✗	✗	✓	✓	Jun '25
POLARIS-1.7B	✓	✓	✗	✗	✓	✓	✓	✓	✓	Jul '25
ProRL-V2	✓	✓	✗	✓	✓	✗	✗	✓	✓	Aug '25
QuestA-Nemotron	✗	✗	✗	✗	✗	✗	✗	✓	✓	Sep '25
BroRL	✓	✓	✗	✓	✓	✗	✗	✓	✓	Oct '25
JustRL-DeepSeek	✓	✗	✗	✗	✗	✗	✗	✗	✗	Nov '25
JustRL-Nemotron	✓	✗	✗	✗	✗	✗	✗	✗	✗	Nov '25

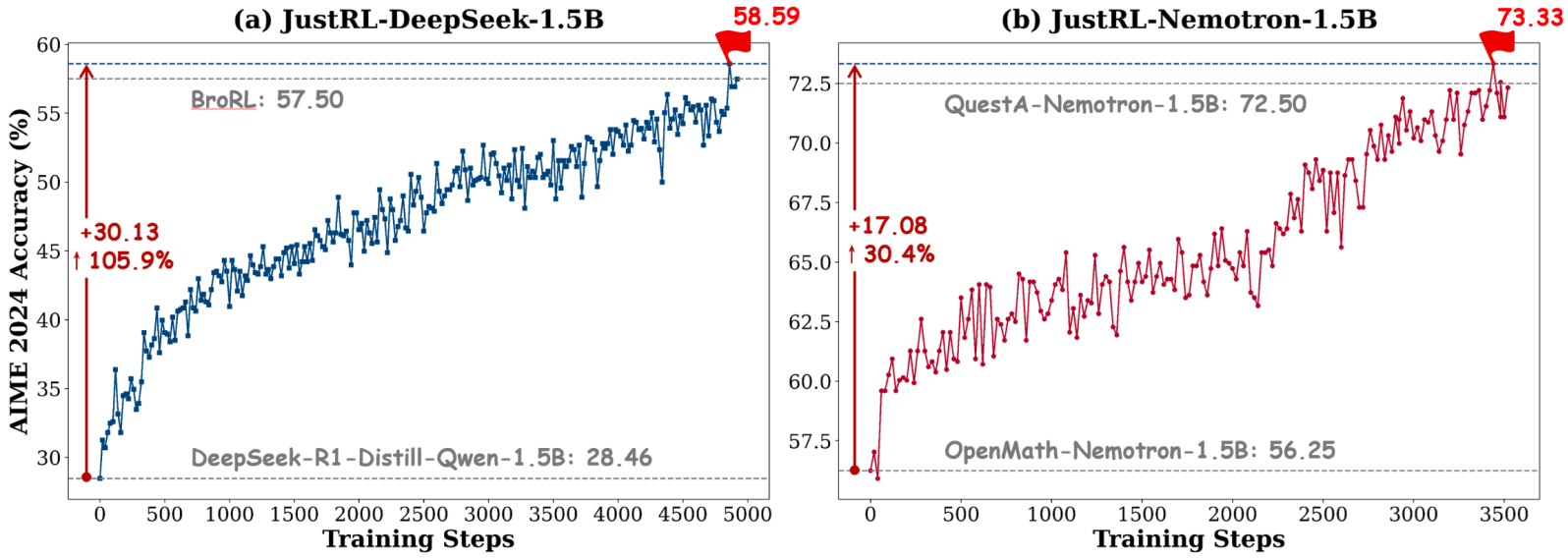
Table 1 | Comparison of RL techniques used in recent small language models for mathematical reasoning. Model names are colored by backbone: DeepSeek-R1-Distill-Qwen-1.5B, Qwen3-1.7B, OpenMath-Nemotron-1.5B. We use the following abbreviations for RL techniques: EC=Entropy Control, THP=Tune Hyperparameters, TTP=Tune Training Prompt, RKL=Reset KL Reference, LC=Length Control, AT=Adaptive Temperature, RR=Rollout Rescue, DS=Dynamic Sampling, ST=Split Training Stages. Our models (JustRL-DeepSeek and JustRL-Nemotron) use only entropy control, achieving competitive performance with minimal complexity.

JustRL: Settings and Results

Core Algorithm : GRPO

What kept simple
- Single-stage training
- Fixed hyperparameters
- Standard data
- Basic Prompting
- Length Control

One technique used
- Clip-higher (entropy control)



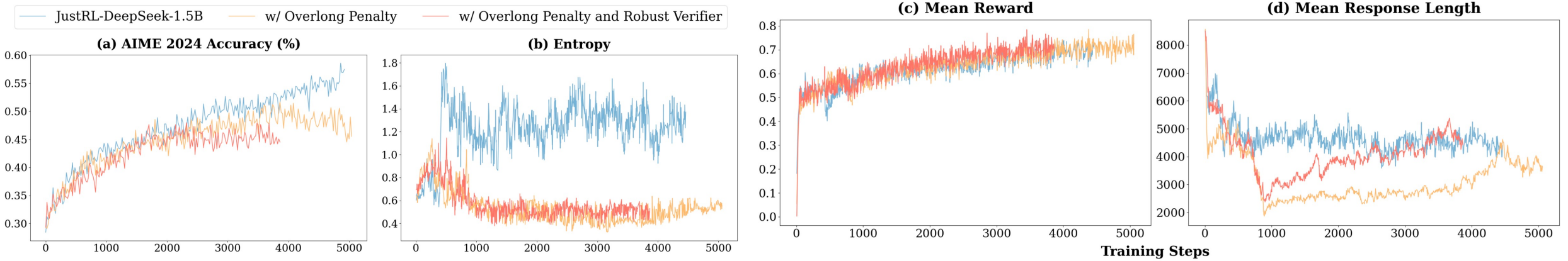
JustRL: Results and Comparisons

Model	AIME24	AIME25	AMC23	MATH	Minerva	Olympiad	HMMT	BRUMO	CMIMC	Avg
Backbone	29.90	22.40	63.82	84.90	34.65	45.95	13.44	30.94	12.89	37.65
DeepScaleR-1.5B	40.21	28.65	73.83	89.30	39.34	52.79	18.96	40.00	21.00	44.88
ProRL-V2	51.87	35.73	88.75	92.00	49.03	67.84	19.38	47.29	**25.86**	53.08
BroRL*	**57.50**	36.88	–	**92.14**	49.08	61.54	–	–	–	–
JustRL-DeepSeek	52.60	**38.75**	**91.02**	91.65	**51.47**	**67.99**	**21.98**	**52.71**	25.63	**54.87**

Table 3 | Results on DeepSeek-R1-Distill-Qwen-1.5B backbone. All scores except MATH-500, Minerva, and OlympiadBench use @32 sampling; those three use @4. *BroRL results are officially reported but models not released; some benchmarks unavailable.

Model	AIME24	AIME25	AMC23	MATH	Minerva	Olympiad	HMMT	BRUMO	CMIMC	Avg
Backbone	58.75	48.44	90.55	92.40	26.93	71.70	30.10	61.67	30.08	56.74
QuestA	**71.56**	62.08	93.44	92.95	**32.08**	72.28	**40.94**	**67.50**	41.48	63.81
JustRL-Nemotron	69.69	**62.92**	**96.02**	**94.15**	30.24	**76.59**	40.63	66.88	**41.72**	**64.32**

Table 5 | Results on OpenMath-Nemotron-1.5B backbone. All scores except MATH-500, Minerva, and OlympiadBench use @32 sampling; those three use @4.



JustRL: Limitation

1. **Limited to mathematical reasoning tasks**
2. **Cannot definitively isolate which specific components are most critical to our success.**
3. Their compute budget may still be prohibitive for resource-constrained researchers
4. Not explored whether their approach maintains advantages when pushed to even longer training horizons or whether additional techniques might become necessary at scale