ACL 2026 SAN DIEGO JULY 2-7

香 港 大 學
THE UNIVERSITY OF HONG KONG

清華大學
Tsinghua University

Revisiting Model Interpolation for Efficient Reasoning

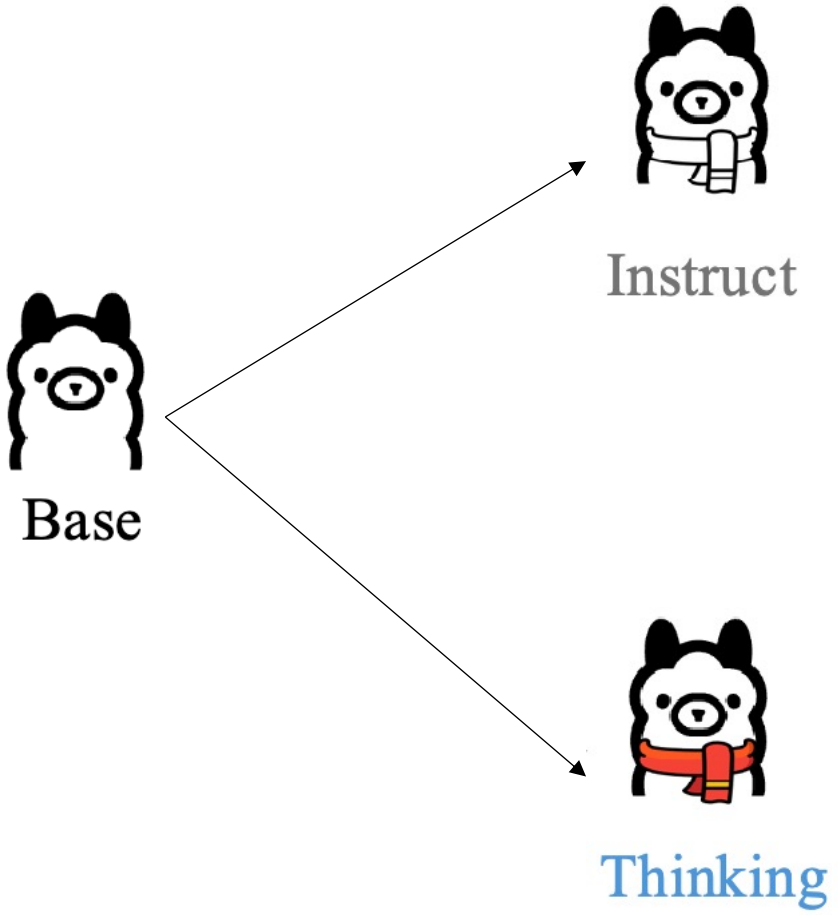
Taiqiang Wu[1], Runming Yang[1], Tao Liu[2], Jiahao Wang[1], Ngai Wong[1]

[1]The University of Hong Kong, [2]Tsinghua University

Paper

Code

Preliminary



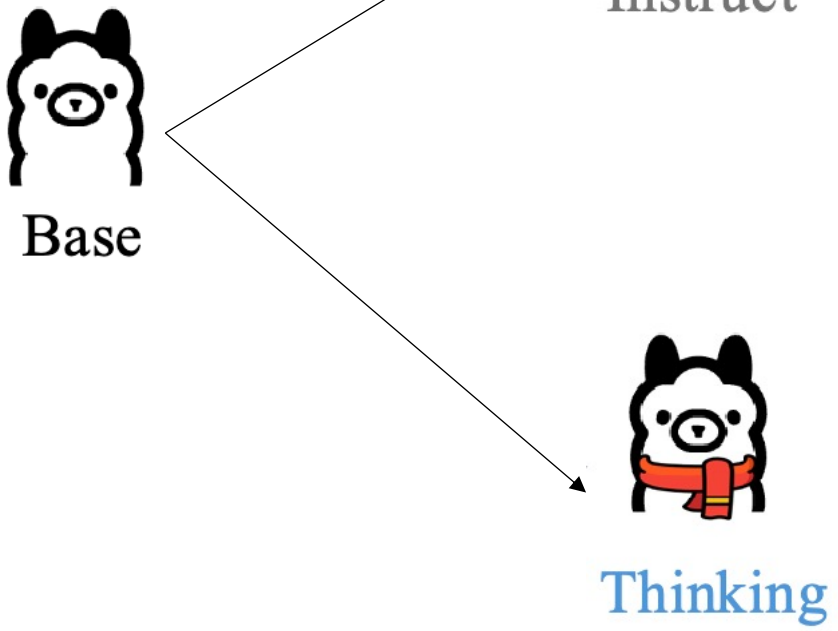
Fast thinking: shorter CoTs

with weaker performance

Slow thinking: longer CoTs

with stronger performance

Preliminary



Fast thinking: shorter CoTs

with weaker performance

Same structure

Distinct behaviors

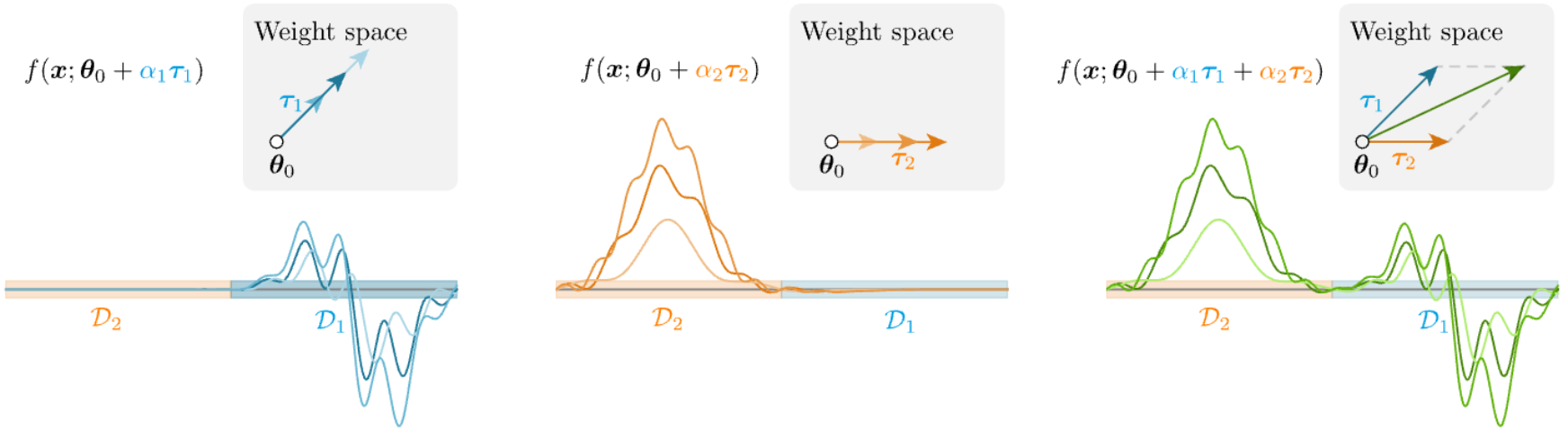
Slow thinking: longer CoTs

with stronger performance

Can we merge the weights?

& What and Why

Model Merging: merge the task vectors to merge capability[1]



Task vector is a small disturbance compared to pre-trained weights

-> Merging occurs within **a local neighborhood**[2]

Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models

Linear Mode Connectivity and the Lottery Ticket Hypothesis

Weight Similarity

Are the weights similar for paired

Instruct and Thinking LLMs?

SHADOW-FT: TUNING INSTRUCT MODEL VIATRAINING ON PAIRED BASE MODEL

Weight Similarity

Are the weights similar for paired Instruct and Thinking LLMs?

Weight similarity[1]:

$$\sigma = \frac{\sum |W_B - W_I|}{\sum |W_B| + \sum |W_I|},$$

Models		σ
Qwen3-4B-Base	Qwen3-4B	0.0326
Qwen3-4B-Base	Qwen3-4B-Instruct	0.0562
Qwen3-4B-Base	Qwen3-4B-Thinking	0.0638
Qwen3-4B	Qwen3-4B-Instruct	0.0589
Qwen3-4B	Qwen3-4B-Thinking	0.0633
Qwen3-4B-Instruct	Qwen3-4B-Thinking	0.0269

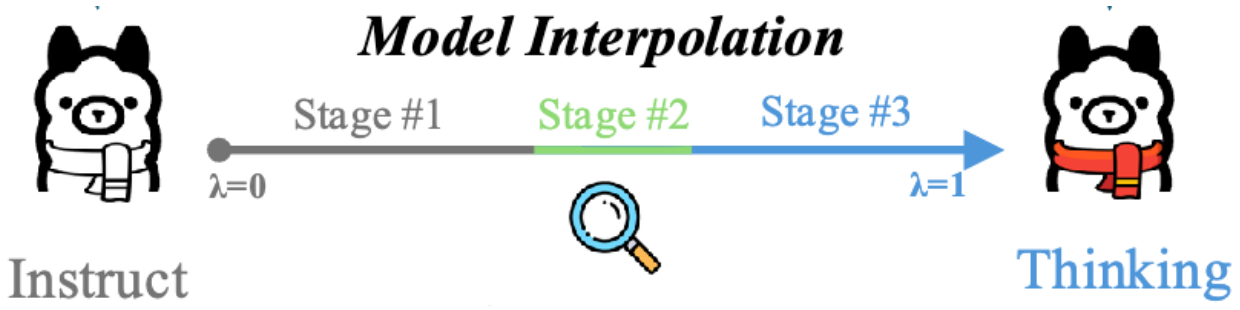
Table 1: Weight similarity σ (Wu et al., 2025b) on paired models from Qwen3 series. We omit the suffix `-2507` for simplicity. The smaller σ, the more similar.

Yeap, there are quite similar in weights

🤔 Guess: using almost same training prompts

SHADOW-FT: TUNING INSTRUCT MODEL VIATRAINING ON PAIRED BASE MODEL

Linear Interpolation



$$\lambda\Theta^{(\text{Thi})} + (1-\lambda)\Theta^{(\text{Ins})} = \lambda(TV^{(\text{Thi})} + \Theta^{(\text{Base})}) + (1-\lambda)(TV^{(\text{Ins})} + \Theta^{(\text{Base})}) = \Theta^{(\text{Base})} + \lambda TV^{(\text{Thi})} + (1-\lambda)TV^{(\text{Ins})}. \quad (7)$$

The $\Theta^{(\text{Base})}$ can be an *arbitrary* model. This deriva-

👉 Model Interpolation (MI) is equivalent to performing Task Arithmetic on the Thinking and Instruct task vectors based on an **arbitrary base model** with scaling factors of λ and $(1-\lambda)$.

Evaluation Protocol

Paired LLMs:

Qwen3-4B-Instruct-2507 & Qwen3-4B-Thinking-2507

Qwen3-30B-A3B-Instruct-2507 & Qwen3-30B-A3B-Thinking-2507

Llama-3.1-8B-Instruct & Deepseek-R1-Distill-Llama-8B

Benchmark:

AIME'25 (roll 64, 80k), IFEval/GPQA-Diamond (roll 8, 64k)

Metric:

Mean@k, Pass@k, Vote@k, Ouput Token Length, Think Ratio (ratio for </*think*>)

Performance

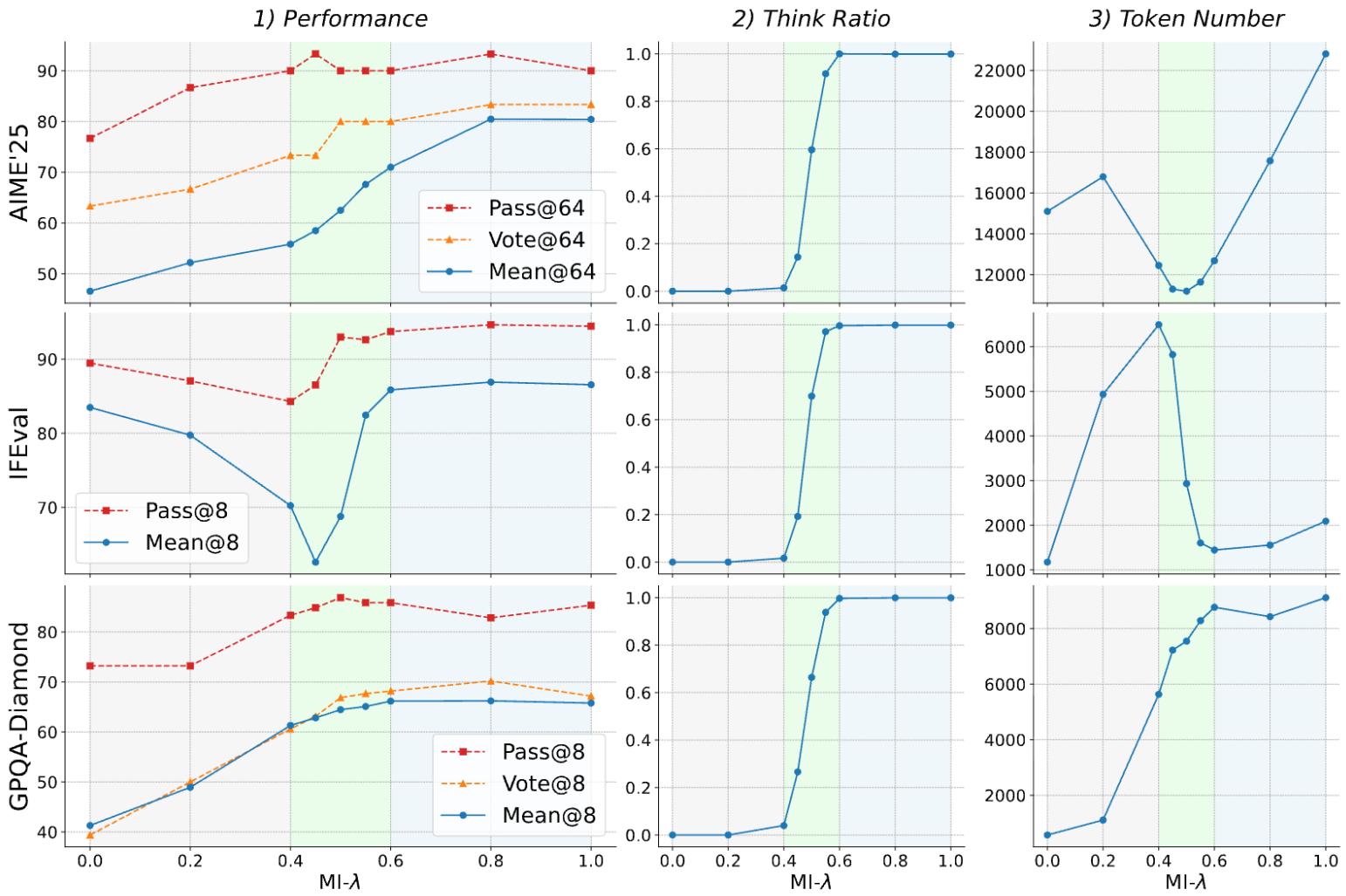


Figure 3: The performance dynamics of model interpolation (MI) on `Qwen3-4B-Instruct-2507` and `Qwen3-4B-Thinking-2507`. The dynamics follow a three-stage evolutionary paradigm colored in grey, green, and blue. λ denotes the interpolation coefficient ranging from 0 to 1. Please refer to Appendix B and Appendix C

Three Stage for λ∈[0,1]

Stage I: 0~0.4

- Think Ratio ~ 0
- Performance better
- Token length larger

(Expect for IFEval)

-> generating longer outputs without adopting an explicit thinking process

Performance

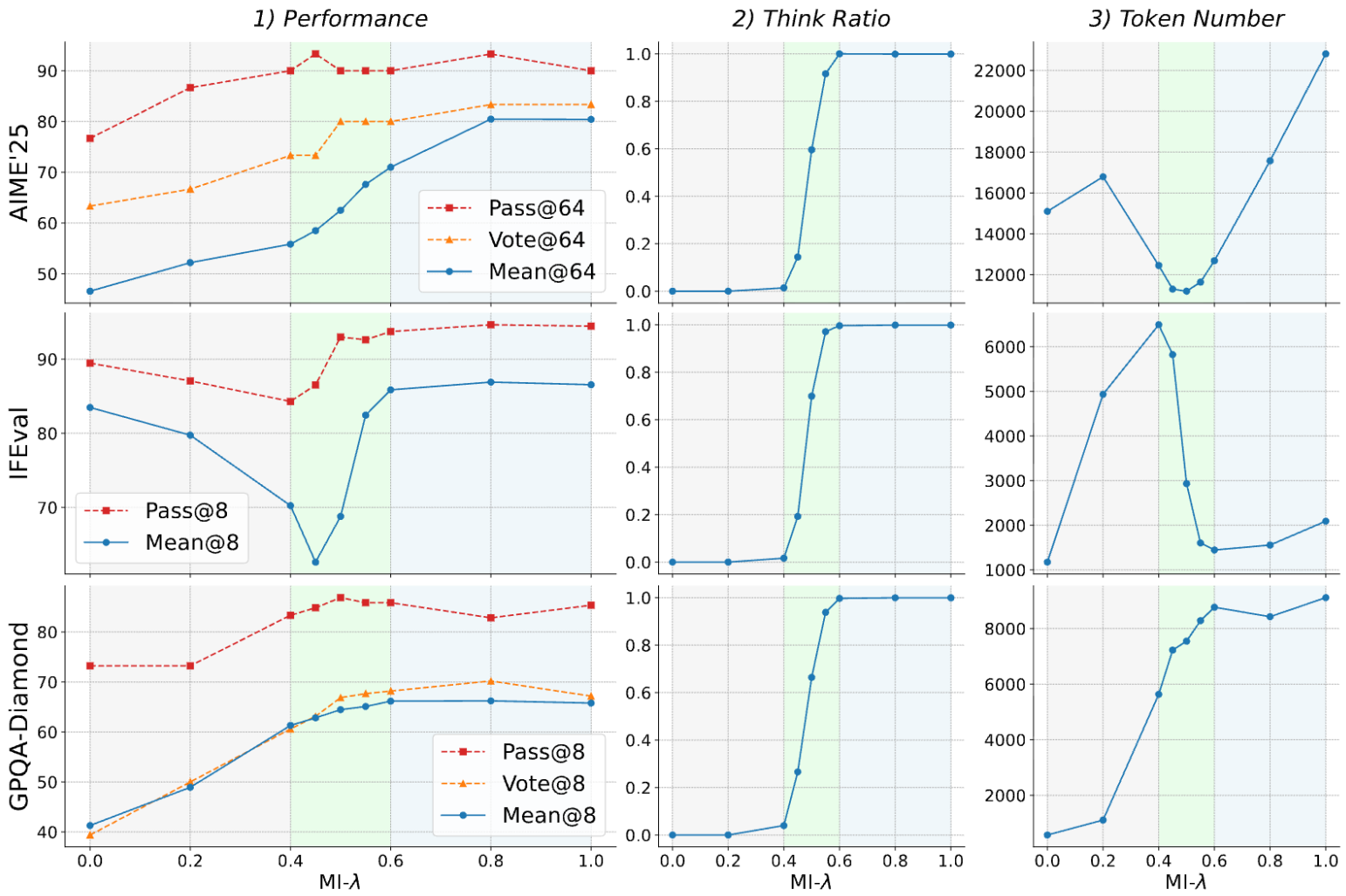


Figure 3: The performance dynamics of model interpolation (MI) on `Qwen3-4B-Instruct-2507` and `Qwen3-4B-Thinking-2507`. The dynamics follow a three-stage evolutionary paradigm colored in grey, green, and blue. λ denotes the interpolation coefficient ranging from 0 to 1. Please refer to Appendix B and Appendix C

Three Stage for λ∈[0,1]

Stage II: 0.4~0.6

- Think Ratio 0->1 rapidly
- Mean@k ↑ ↑ Pass@k ↑
- Token length change rapidly

-> reasoning pattern following Thinking models **rapidly** emerges

Performance

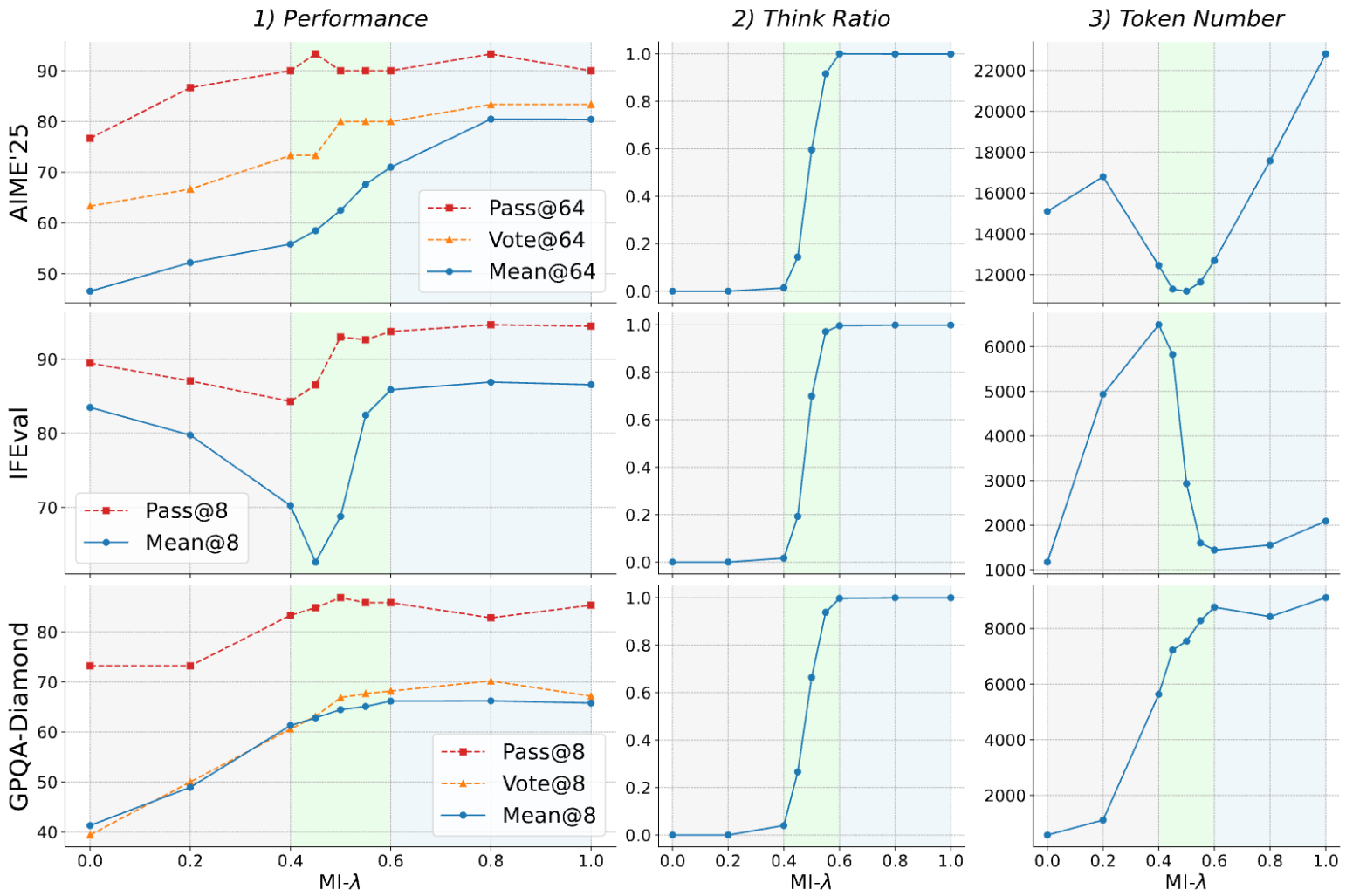


Figure 3: The performance dynamics of model interpolation (MI) on `Qwen3-4B-Instruct-2507` and `Qwen3-4B-Thinking-2507`. The dynamics follow a three-stage evolutionary paradigm colored in grey, green, and blue. λ denotes the interpolation coefficient ranging from 0 to 1. Please refer to Appendix B and Appendix C

Three Stage for λ∈[0,1]

Stage III: 0.6~1

- Think Ratio ~1
- Mean@k & Pass@k ~
- Token length larger

-> converges to the pure Thinking model, with continuously increasing token and slight change performance

Performance

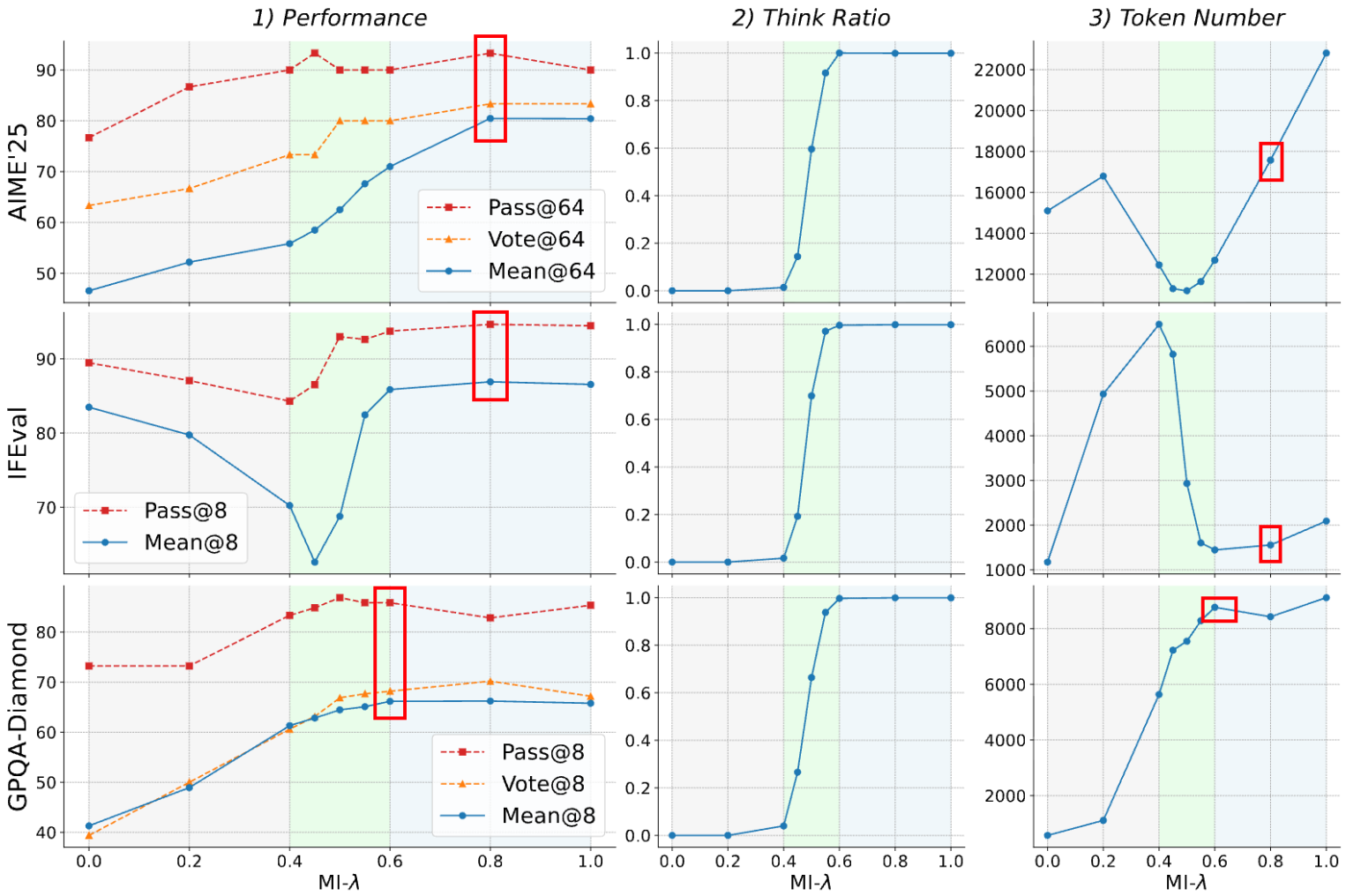


Figure 3: The performance dynamics of model interpolation (MI) on `Qwen3-4B-Instruct-2507` and `Qwen3-4B-Thinking-2507`. The dynamics follow a three-stage evolutionary paradigm colored in grey, green, and blue. λ denotes the interpolation coefficient ranging from 0 to 1. Please refer to Appendix B and Appendix C

One interesting thing:

We can always find a sweet spot λ^* with higher Mean@64, comparable Pass@64 and less token compared to Thinking model. But λ^* conditioned on LLM and benchmark.

Performance

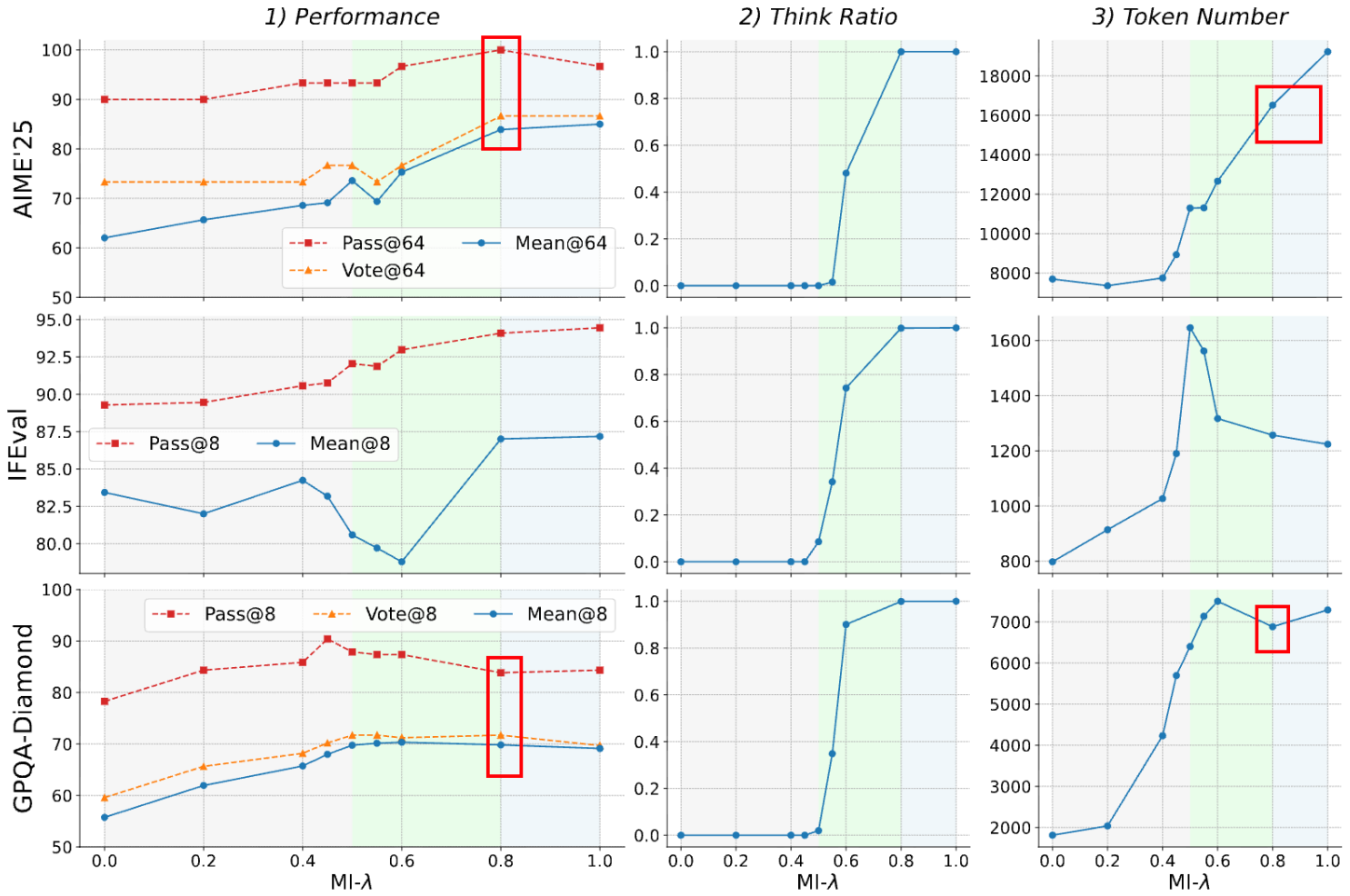


Figure 6: The performance dynamics of model interpolation (MI) on `Qwen3-30B-A3B-Instruct-2507` and `Qwen3-30B-A3B-Thinking-2507`. The dynamics follow a three-stage evolutionary paradigm, while the division range of λ is different.

For Qwen3-30B-A3B

Three stage & sweet spot

Performance

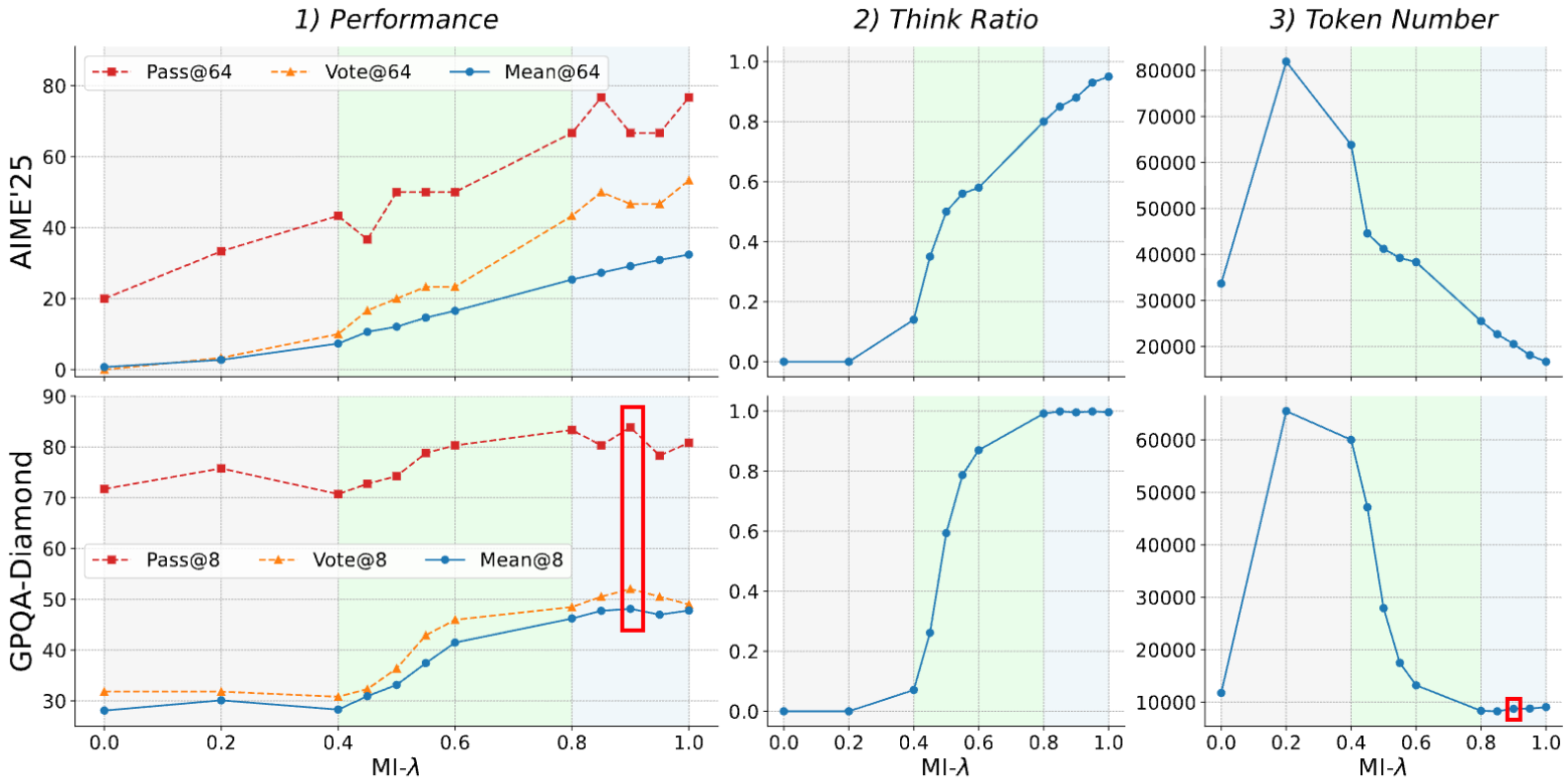


Figure 7: The performance dynamics of model interpolation (MI) on `Llama-3.1-8B-Instruct` and `DeepSeek-R1-Distill-Llama-8B`. The dynamics follow a three-stage evolutionary paradigm, while the division range of λ is different.

For Llama-3.1-8B

Three stage & sweet spot

*There is no official thinking model so we use DS-distilled model instead

Ablation Study

Instruct

Top-p 0.8

Temperature 0.7

≠

Top-p 0.95

Temperature 0.6

Thinking

What is best Top-p and Temperature for merged LLM?

Ablation Study

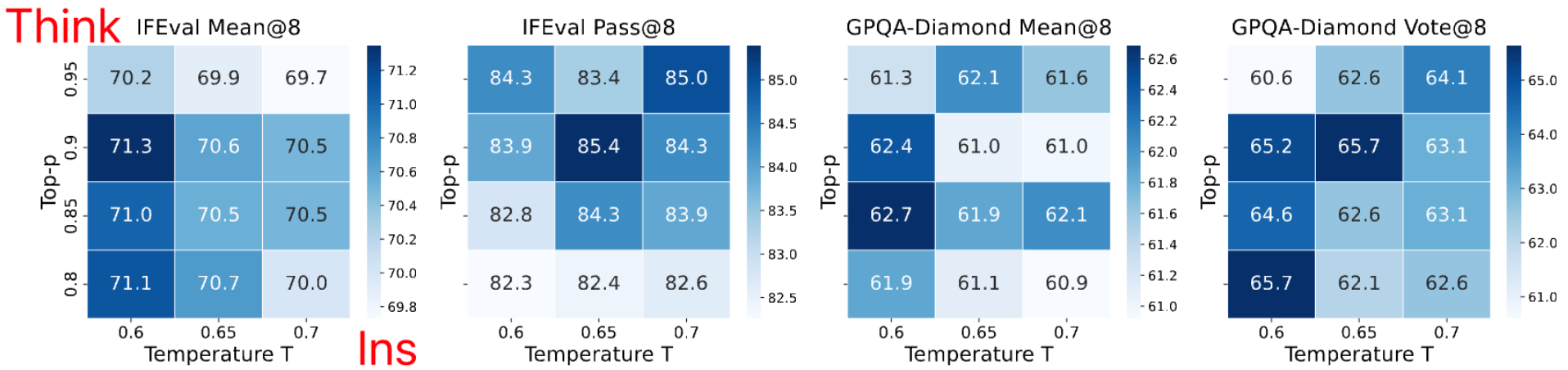


Figure 4: Performance of MI-0.4 on IFEval and GPQA-Diamond under different decoding strategies on Qwen3-4B. We search for the temperature T and Top-p.

- Merged LLM is quite robust for hyper-parameter
- Setting on the Thinking model is a slightly better choice

Ablation Study

Layers to merge

Model	Layers	Mean@64	Pass@8	Pass@32	Pass@64	Vote@64	Token #N	Think #R
Instruct	-	46.57	68.44	74.09	76.67	63.33	15097	0.00
Thinking	-	80.42	89.43	90.00	90.00	83.33	22813	99.95
MI-0.8	[0, 35]	80.47	90.33	91.67	93.30	83.33	17574	99.95
	[0, 11]	42.50	71.25	77.96	80.00	50.00	32151	0.00
	[12, 23]	54.69	78.05	85.53	86.67	63.30	20679	10.57
	[24, 35]	51.35	72.92	78.10	80.00	63.30	14987	31.20
	[0, 23]	59.06	83.17	89.60	93.33	70.00	18044	48.13
	[12, 35]	69.48	85.61	89.17	90.00	76.66	13159	100.00

Table 3: Ablation on different layers to apply model interpolation. **Layers** denote the position to apply interpolation. There are 36 layers in total. We can find that the last two-thirds of the model layers are vital for the thinking pattern.

- The last two-thirds of the model layers are vital for the thinking pattern.

Ablation Study

Module to merge

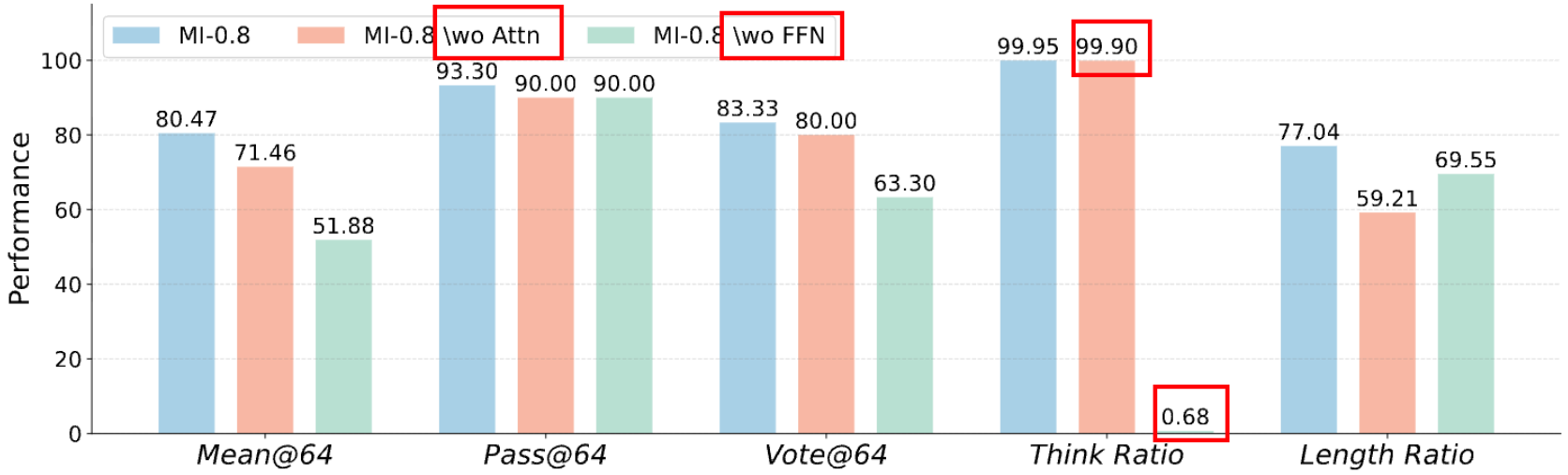


Figure 5: Ablation on modules to apply model interpolation. Attn denotes the MHA sub-layers and FFN for FFN sublayers. We report the results on AIME'25. Length Ratio denotes the ratio to the Thinking model.

- FFN modules from the Thinking model are the primary drivers for the pattern of long CoT reasoning.

Future Work

For weight similarity during LLM post-train:

[Slides] 

[Video] 

For future work:

- Simultaneous interpolation of three or more specialist
- Extrapolation: $\lambda<0$ or $\lambda>1$