



Smart Picks in the Dark: Towards Efficient RLVR for Reasoning via Tracing Metacognitive Pivots

Guangcheng Zhu, Shenzhi Yang, Haobo Wang, Xing Zheng, Yingfan Ma
Xuening Feng, Zhongqi Chen, Bowen Song, Weiqiang Wang, Gang Chen

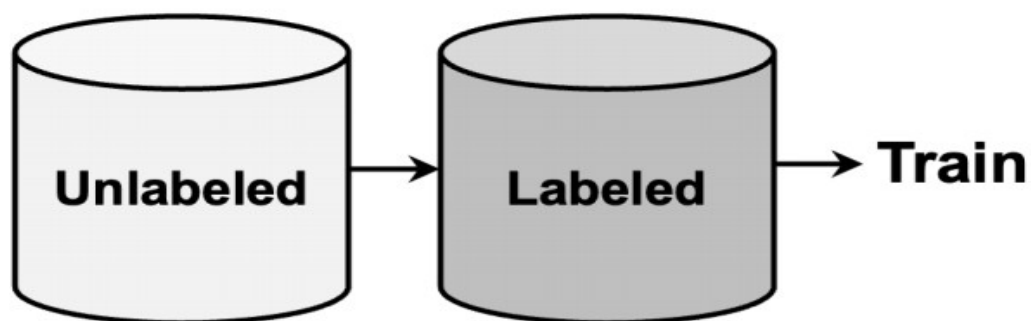
Background



■ Two sources of cost in RLVR

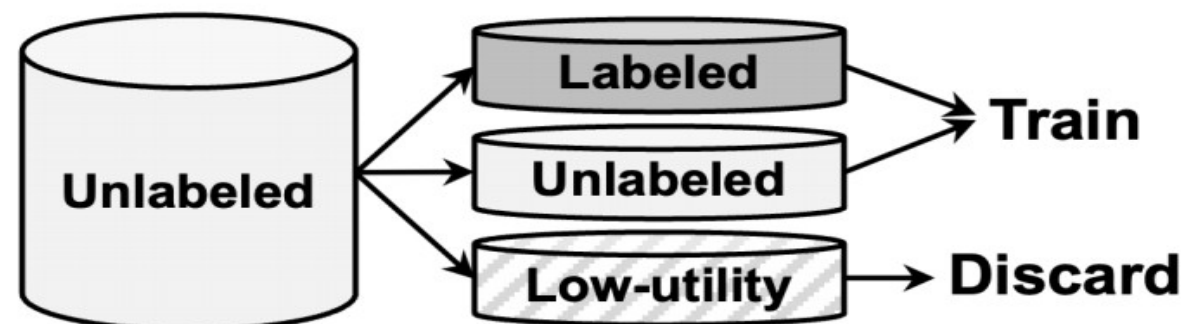
- **Annotation cost:** Reliable labels may require expert input.
- **Training cost:** Repeated long rollouts consume compute.

Standard RLVR



✂ Annotation: **100%**
⚡ Training Steps: **440**
🎯 Performance: **49.2 / 64.3**

PivotTrace (Ours)



✂ Annotation: **29.3%** → **Fewer**
⚡ Training Steps: **160** → **Faster**
🎯 Performance: **49.5 / 64.9** → **Finer**

■ Data selection usually relies on labeled answers

- Accuracy reflects model mastery, but requires reference answers.
- This spends the annotation budget before selection even begins.

■ Self-generated rewards are useful only when reliable

- Unreliable self-feedback can reinforce errors during training.
- Annotation should target questions whose self-feedback we cannot trust.

■ Pick in the Dark: select questions before labeling

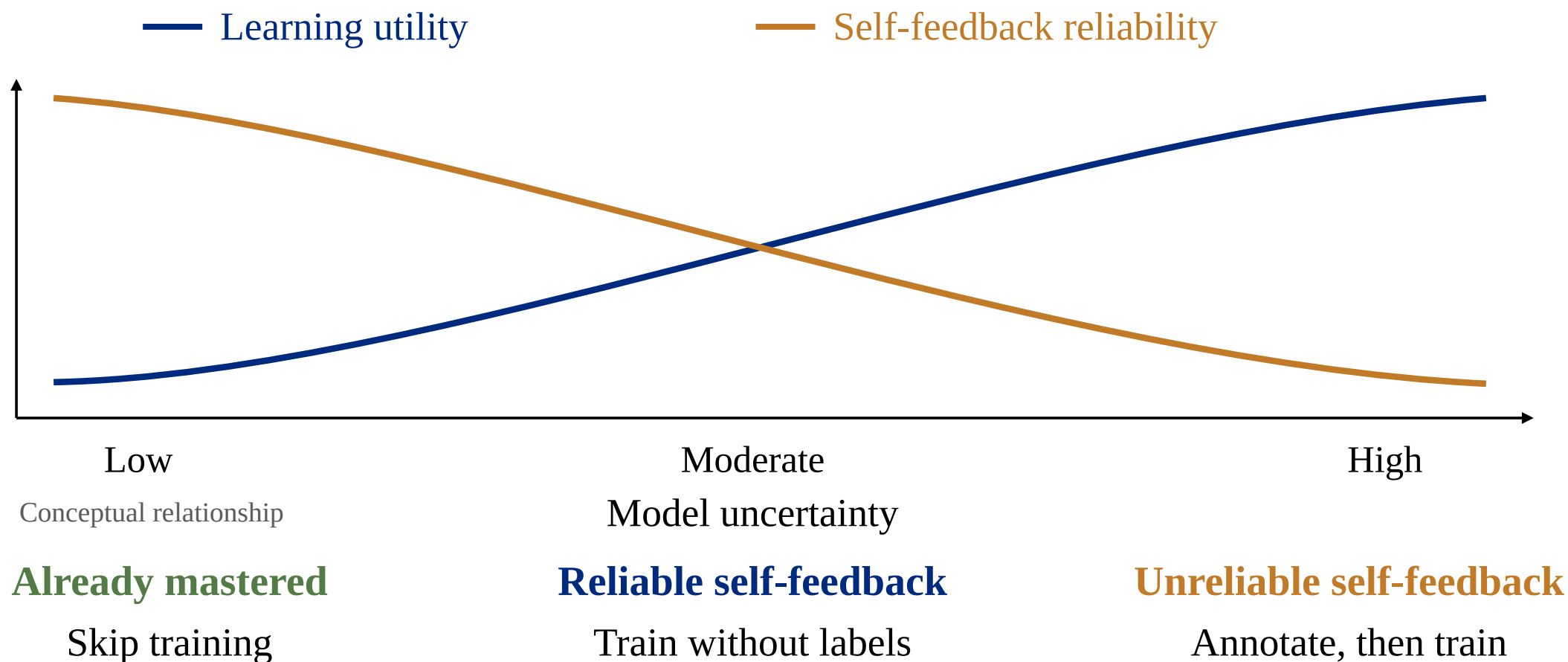
- This requires estimating model uncertainty without labeled answers.

Motivation



■ Model uncertainty guides what to train on and what to label

- Learning utility and self-feedback reliability move in opposite directions.

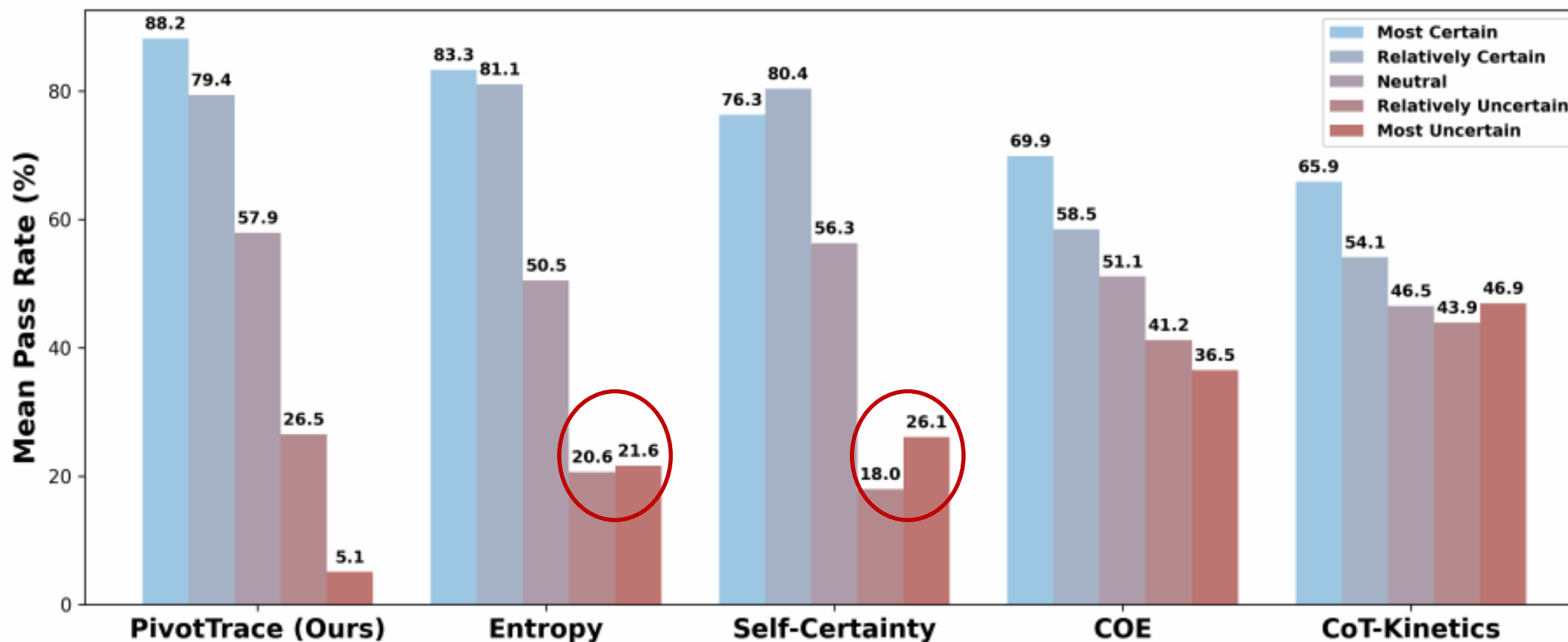


Motivation



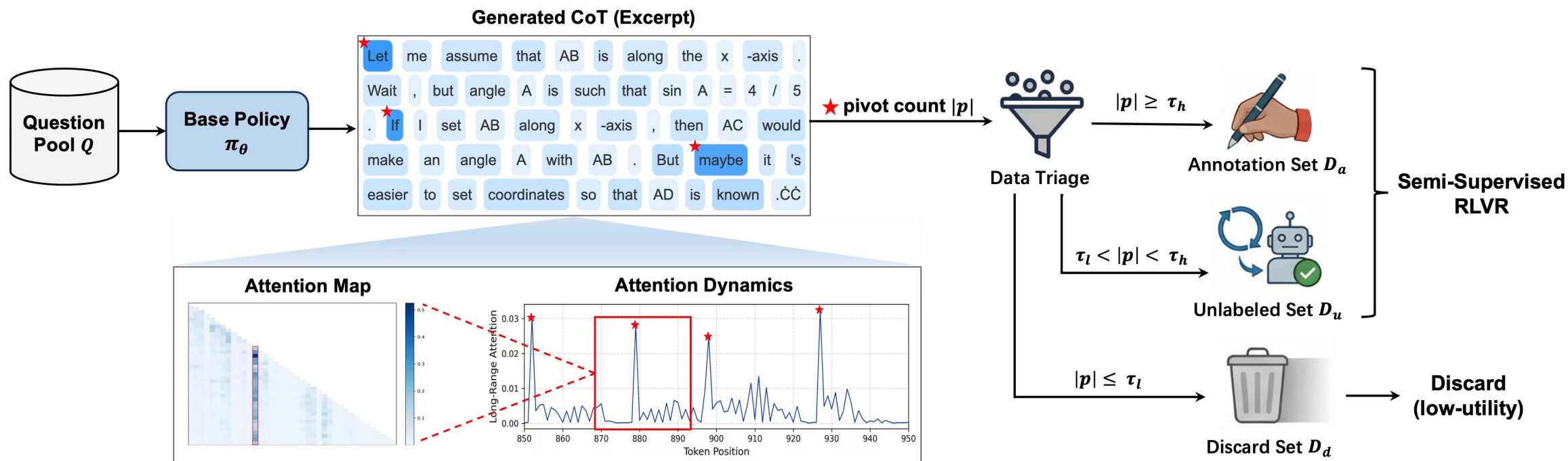
Existing uncertainty scores do not reliably rank model mastery

- **Weak separation:** CoE and CoT-Kinetics.
- **Misordered certainty levels:** entropy and self-certainty.



■ PivotTrace: uncertainty-guided data selection

- Trace metacognitive pivots to estimate uncertainty without labels.
- Calibrate selection cutoffs, then route questions to training or annotation.

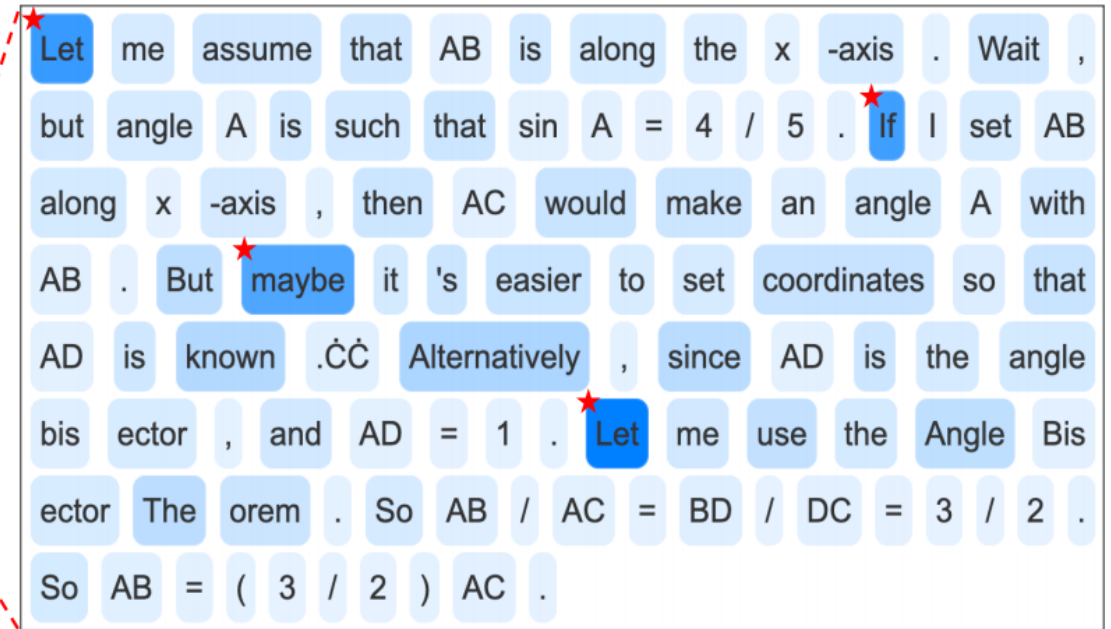
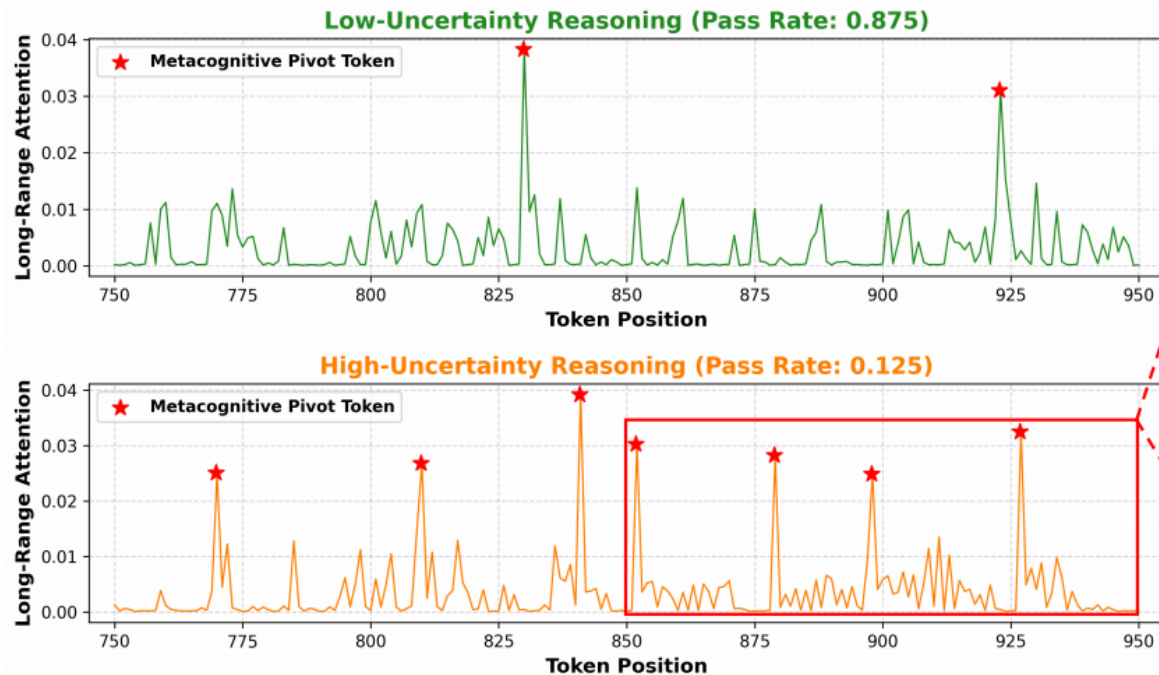


Metacognitive Pivots



■ Like humans, uncertain models may revise earlier reasoning

- These shifts, called **metacognitive pivots**, attract sustained attention from later tokens.
- **More pivots indicate greater model uncertainty.**



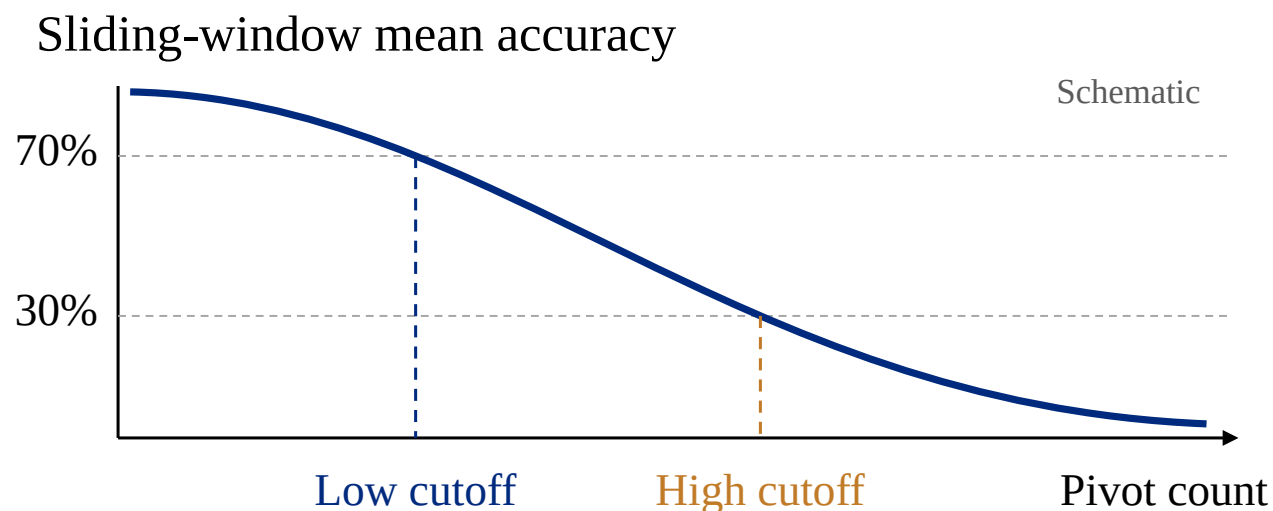
Token-Level Visualization of Long-Range Attention Distribution

Attention peaks reveal key reasoning shifts

Uncertainty-Guided Data Selection

■ Calibrate two cutoffs using a small labeled probe set

- Rank by pivot count, sample 100 probes, and track local mean accuracy.



Low cutoff

First window below
70% accuracy

High cutoff

First window below
30% accuracy

■ Apply the cutoffs to the full unlabeled question pool

Low uncertainty



Discard

Low learning utility

Moderate uncertainty



Train without labels

Majority-vote rewards

High uncertainty



Annotate, then train

Verified rewards

Experiments



■ Better accuracy with fewer labels and less training data

➤ **29.3% labels** and **57.9% training data** outperform full supervision.

Methods	In-Domain Performance						Out-of-Domain Performance			
	AIME 24/25	AMC	MATH-500	Minerva	Olympiad	Avg.	ARC-c	GPQA*	MMLU-Pro	Avg.
Semi-Supervised Training on the Selected Subset ($\rho_t \approx 57.9\%$) with $\rho_a \approx 29.3\%$ Annotated										
Random	23.1/23.4	56.6	85.2	42.7	47.5	46.4	88.2	33.0	62.4	61.2
Consistency [†]	25.9/22.1	59.9	86.2	43.5	50.0	47.9	86.1	33.2	62.7	60.7
CoE	26.5/23.9	57.0	87.0	43.3	49.7	47.9	92.8	32.8	61.6	62.4
CoT-Kinetics	21.6/24.8	57.3	86.0	43.7	47.6	46.8	88.7	33.7	61.8	61.4
Entropy	26.1/22.1	59.2	86.1	43.7	48.4	47.6	83.3	37.8	62.5	61.2
Self-Certainty	24.2/23.6	58.8	86.3	43.8	48.5	47.5	90.1	35.7	61.7	62.5
PivotTrace (ours)	27.2/25.2	62.4	87.3	44.6	50.1	49.5	93.0	38.3	63.5	64.9
Fully Supervised	26.6/23.8	60.7	87.4	43.8	52.9	49.2	93.1	36.5	63.4	64.3

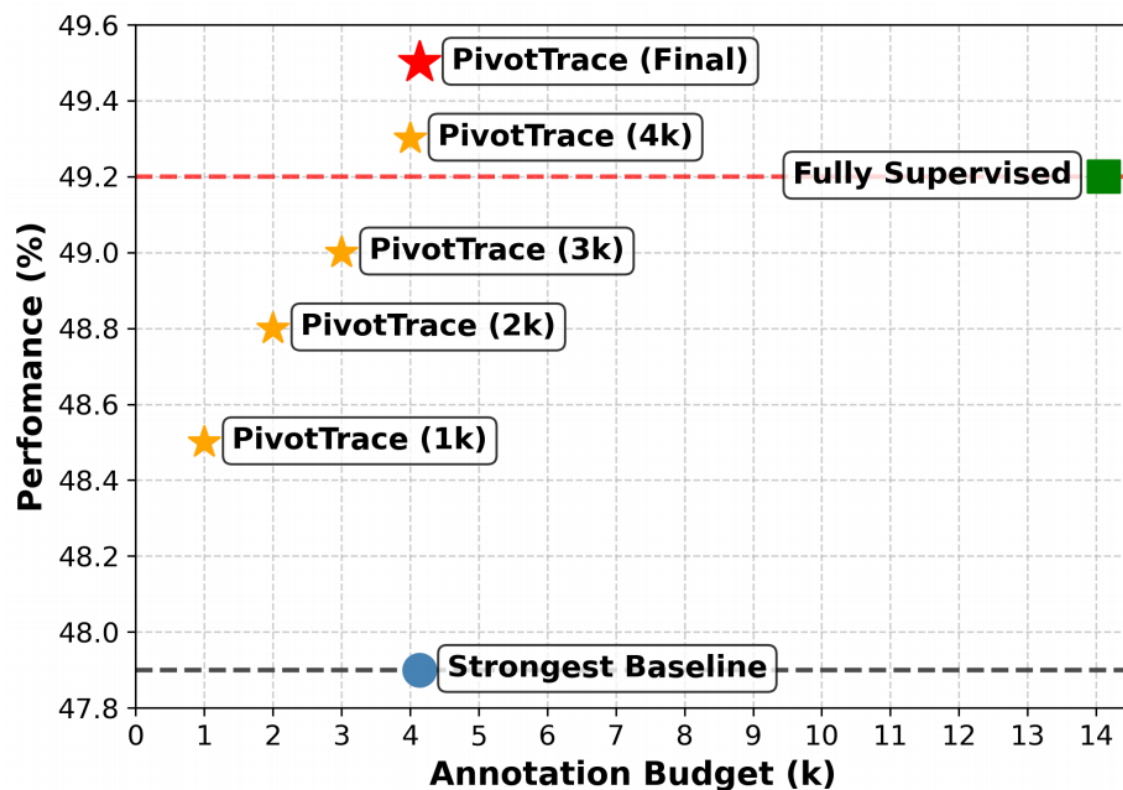
Qwen3-4B-Base on nine in-domain and out-of-domain benchmarks

Experiments



■ Annotation efficiency

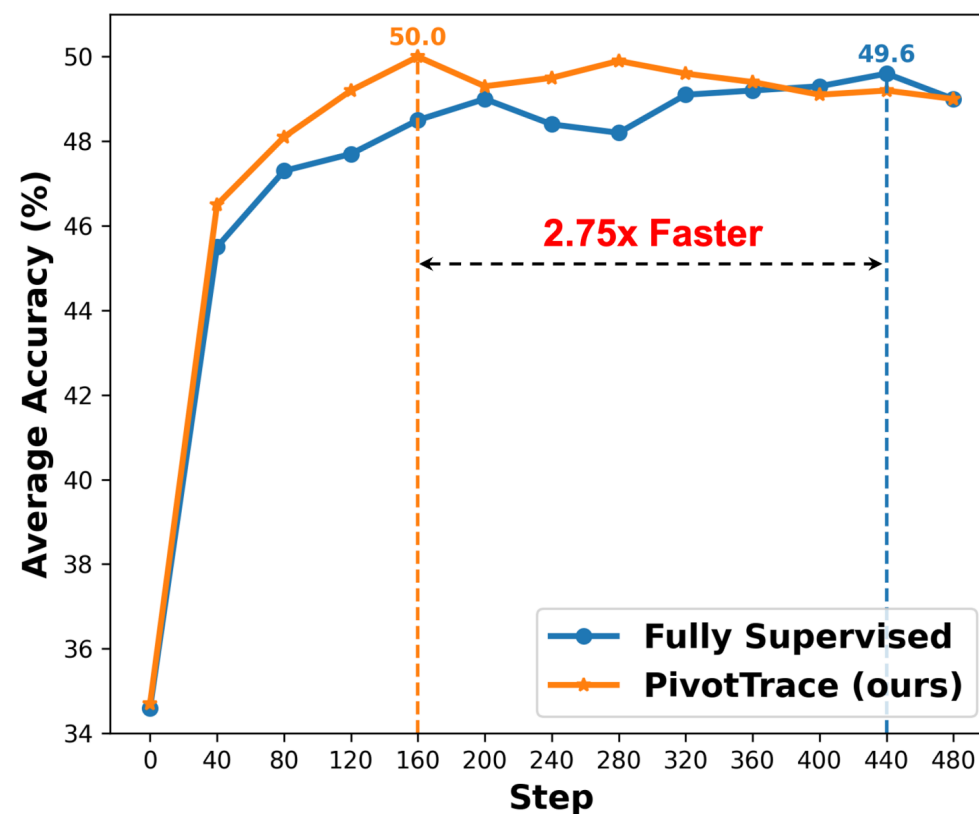
- Beats the strongest baseline with 1k labels.



(a) Performance vs. annotation budget

■ Training efficiency

- Converges 2.75× faster than full supervision.



(b) Convergence in training steps



Thanks for Listening!

Paper: arxiv.org/abs/2606.04503

Code: github.com/gczhu/PivotTrace