

1 Forward Pipeline

1 Input Conversation x



User: Prove that $\sqrt{2}$ is irrational.

2 Student Template c_θ



<|im_start|> user
Prove that...
<|im_end|>
<|im_start|>
assistant<think>

3 Teacher Template c_ϕ



<<s> [INST]
user
Prove that...
[/INST]
assistant

4 Student Rollout



... </think>
...Therefore,
 $\sqrt{2}$ is irrational.
<|im_end|>

5 Compute Student Log Probabilities

6 Append Student Response to Teacher Template and Compute Teacher Log Probabilities

Collect Log Probabilities of Student and Teacher

2 Compute advantages for policy updates

Student Response : Because 2 is prime, if $\sqrt{2} = p/q$, ...<|im_end|>



Student Tokens

2 is prime , if $\sqrt{2} = p / q$,



Teacher Tokens

2 is prime , if $\sqrt{2} = p/q$,

Compute advantages on aligned positions as: $\hat{A}_t = \tilde{\ell}_t^\phi - \log \pi_{\theta_{\text{old}}}(y_t | c_\theta, y_{<t})$ Eq.(7)



Advantages

2 is prime , if $\sqrt{2} = p / q$,

Compute policy loss as: $\mathcal{L}_\theta = -\mathbb{E} \left[\sum_{t=1}^n \min \left(r_t \hat{A}_t, \text{clip}(r_t, 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right]$ Eq.(9)



Student (short-context)
e.g., Qwen3 / Intern-S2



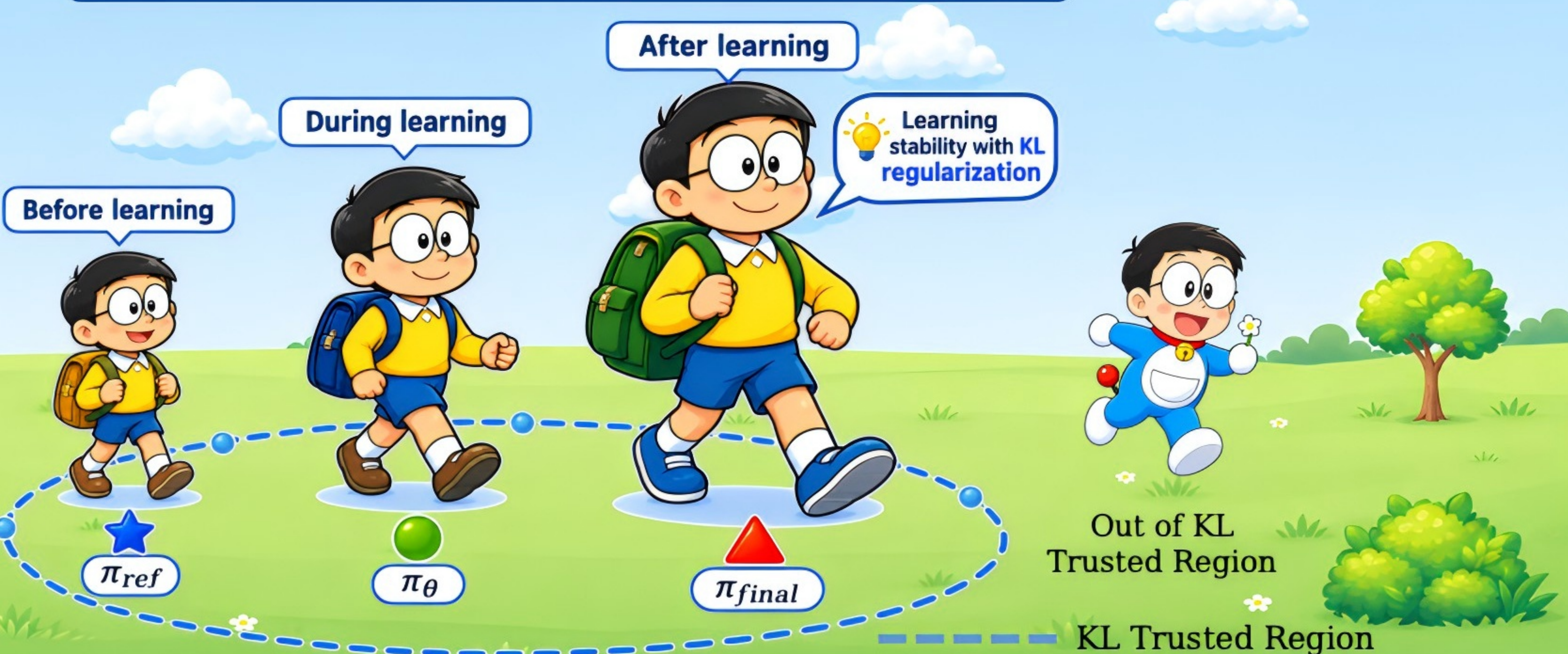
Teacher (long-context)
e.g., SU-01

High Advantages Low



Unmatched tokens
(no teacher signal)

3 KL Constraint: Grow in a Trust Region



4 Benefits



Cross-tokenizer OPD

Align on shared text spans, no full token-level correspondence needed.



Stabilized Training

KL constraint + termination-token masking reduce length explosion



Effective Transfer

Transfers long-context reasoning to students across model families.



Strong Performance

Significant gains on math, proof and science benchmarks.

