



ICLR
International Conference On
Learning Representations

ORAL

Reducing Belief Deviation in Reinforcement Learning for Active Reasoning of LLM Agents

Deyu Zou^{*}, Yongqiang Chen^{*}, Jianxiang Wang, Garry YANG, Mufei Li,
Qing Da, James Cheng, Pan Li, Yu Gong

The Chinese University of Hong Kong
ByteDance
Georgia Institute of Technology

Active Reasoning in Dynamic Environments



Iterative Interaction to solve tasks

Agents trapped in wrong intermediate assumptions

Coding Agent



Endless wrong debugging loop

Computer-use Agent



Repeated failed actions

Dialogue Agent



Off-track conversation

Why do LLM agents get trapped in active reasoning?
Can outcome-based RL automatically fix it?

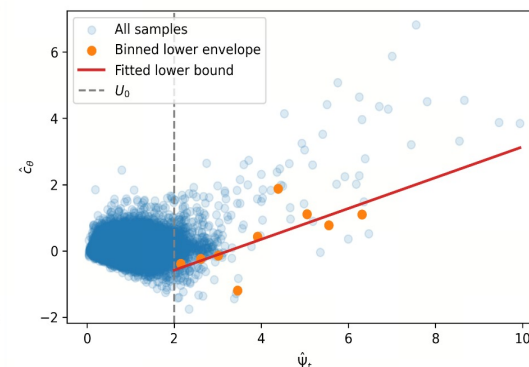
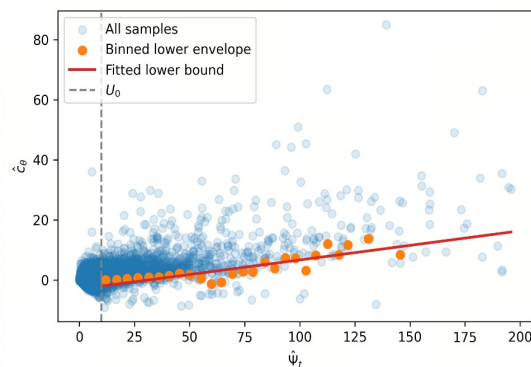
Belief Tracking is Central to Active Reasoning

What is belief tracking?



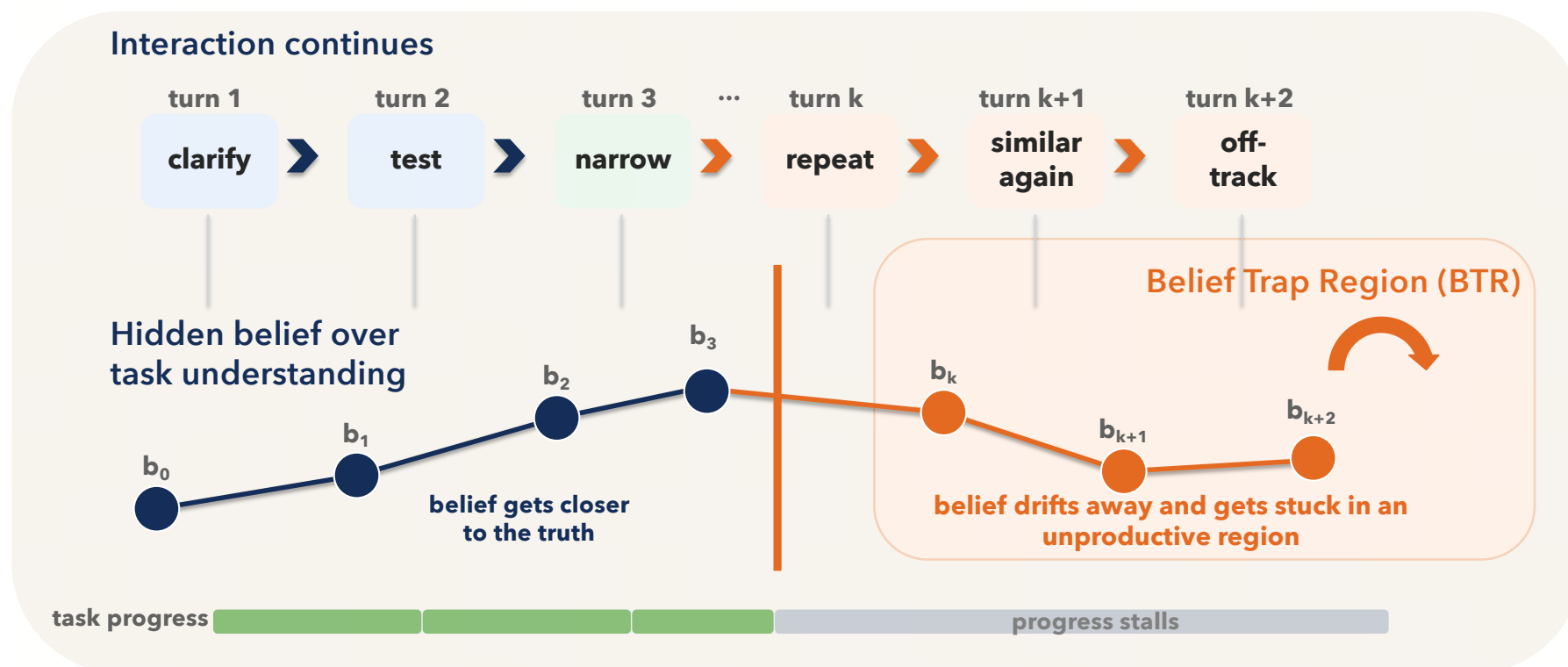
The agent must maintain an internal belief over the latent task state.

LLMs are imperfect belief trackers
(Asmp. 1)



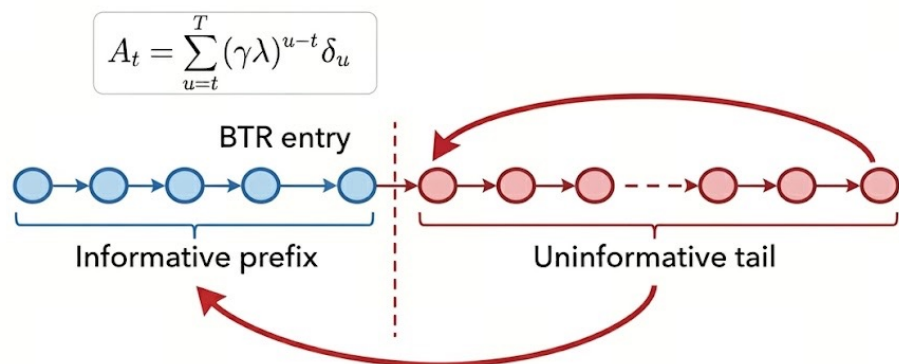
In high-uncertainty regimes, the belief-update error **grows** with the deviation itself

Belief deviation leads to a Belief Trap Region (Thm. 1)

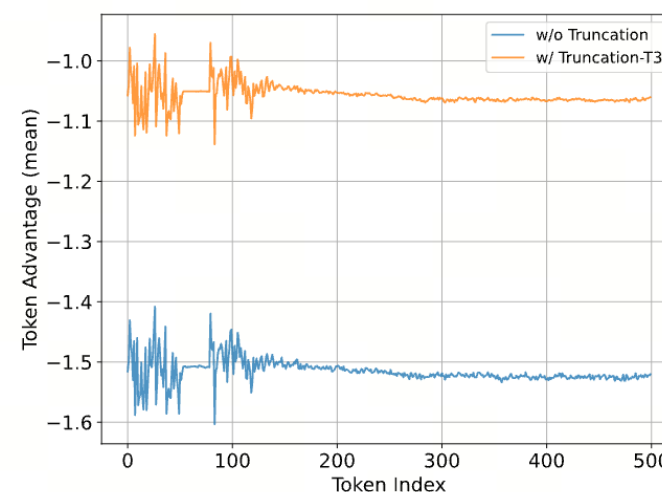


Long trajectory tails may look active on the surface, yet contribute little to actually solving the task once the agent has entered the BTR.

BTR tails corrupt early-token credit (Thm. 2)



Inside BTR, tail TD-errors become systematically harmful, dragging down prefix credit.



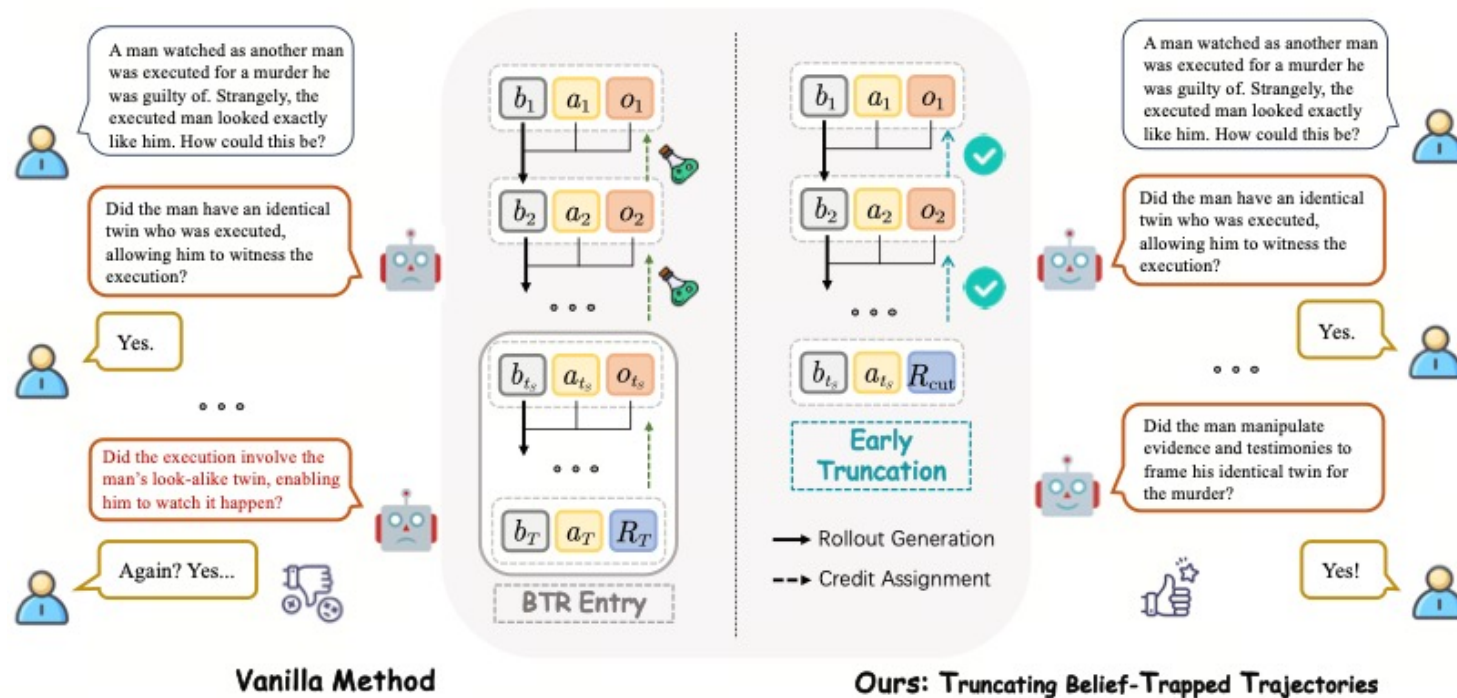
Token-wise mean GAE on failed rollouts

Truncating the uninformative tail preserves cleaner credit for informative prefix

A simple yet powerful principle: Truncating Belief-Trapped Trajectories (T3)

T3: Stop learning from stalled tails (def. 2)

ICLR 2026 ORAL



How to detect BTR entry in practice?

Simple proxy signals are enough (*T3 condition, Def 2*)

ICLR 2026 ORAL

Theory gives the criterion; task traces give the trigger.

GuessNumbers

Cut when the guess contradicts the feasible candidate set.

SituationPuzzles

Cut when questioning becomes *repetitive and semantically similar*.

CircuitDecoding

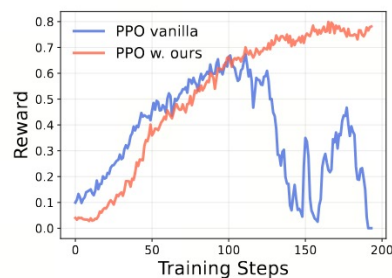
Cut when the candidate circuit set stops shrinking across turns.

PreferenceEstimation / MR

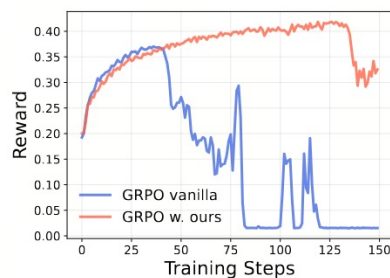
Cut when the preference estimate converges and *does not change meaningfully*.

Capturing more/less task structural knowledge: more general-purpose/task-specific

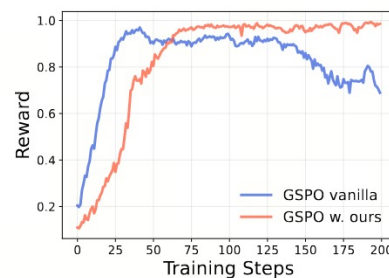
T3 delivers large gains across tasks and algorithms



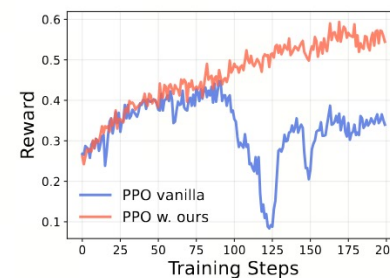
(a) CD (PPO)



(b) SP (GRPO)

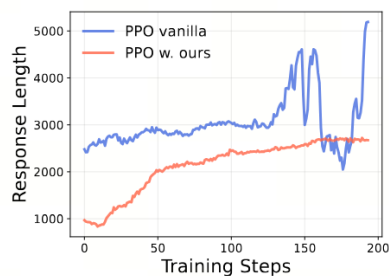


(c) GN (GSPO)

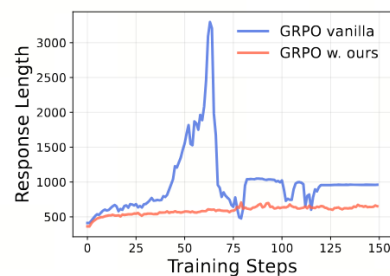


(d) PE (PPO)

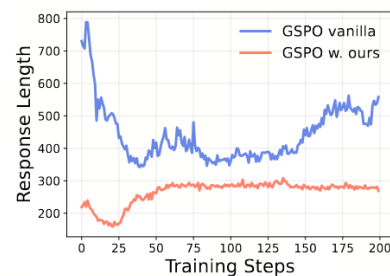
Higher token efficiency



(a) CD (PPO)



(b) SP (GRPO)

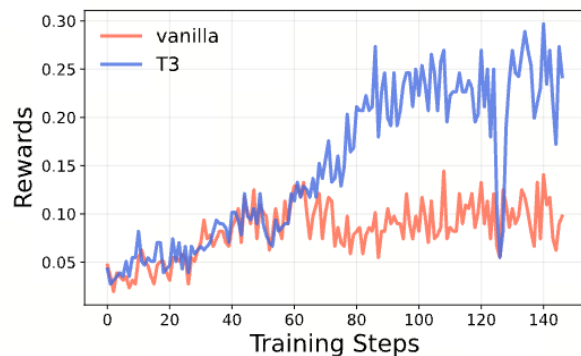


(c) GN (GSPO)

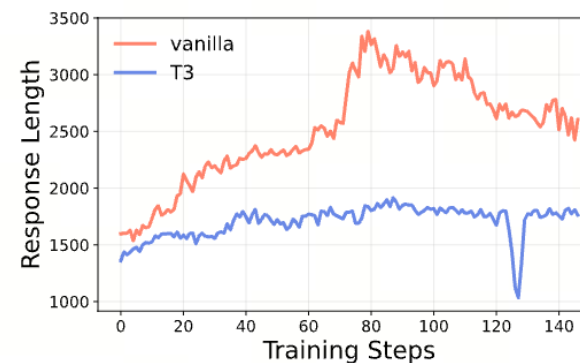


(d) PE (PPO)

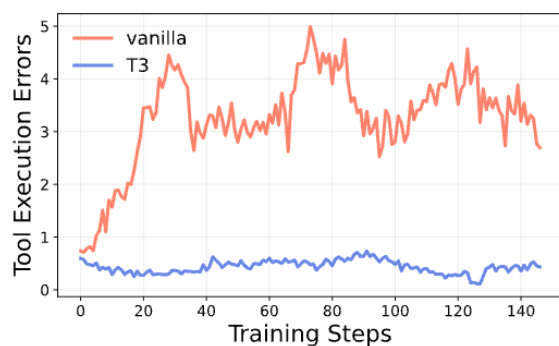
Extending T3 to realistic & popular Agentic scenarios (*tau2-bench*)



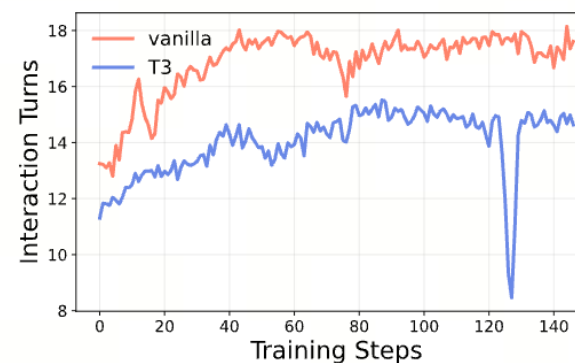
Rewards



Token cost



Tool use error



Interaction turns

Takeaways

ICLR 2026 ORAL

Three Takeaways

1. Belief deviation is a hidden bottleneck in RL for active reasoning.
2. Truncating bad tails protects the credit of good prefixes.
3. T3 is simple, intuitive, and easy to plug into existing RL pipelines.

Future works

1. General-purpose BTR detectors
2. Applicability to deep research, SWE, etc.
3. More mechanism exploration on agentic applications (Welcome to check our new work!)

Welcome to check out our paper
and code!

