Novelty Assessment Report

Paper: $\mathbf{T^3}$: Reducing Belief Deviation in Reinforcement Learning for Active Reasoning
PDF URL: https://openreview.net/pdf?id=r8hzDA3pUY
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Active reasoning requires large language models (LLMs) to interact with external sources and strategically gather information to solve problems. Central to this process is belief tracking: maintaining a coherent understanding of the problem state and the missing information toward the solution. However, due to limited reasoning capabilities, LLM-based agents often suffer from belief deviation: they struggle to correctly model beliefs, lose track of problem states, and fall into uninformative or repetitive actions. Once this happens, errors compound and reinforcement learning (RL) training fails to properly credit the crucial exploratory steps. To address this issue, we propose to track the deviation of model beliefs and develop $\mathbf{T^3}$, a simple yet effective method that detects excessive belief deviation and truncates trajectories during training to remove uninformative tails. By preserving credit for informative prefixes, $\mathbf{T^3}$ systematically improves policy optimization. Across 5 challenging tasks, $\mathbf{T^3}$ consistently enhances training stability, token efficiency, and final performance, achieving up to 30\% gains while cutting rollout tokens by roughly 25\%. These results highlight belief control as a key principle for developing robust and generalizable LLM-based active reasoners.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Reducing Belief Deviation in Reinforcement Learning for Active Reasoning**
A total of **11 papers** were analyzed and organized into a taxonomy with **9 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **LLM-Based Active Reasoning with Belief Tracking**
- **Multiagent Epistemic Planning and Belief Coordination**
- **Active Inference and Belief Dynamics**
- **Hybrid Learning Systems Combining RL and Symbolic Reasoning**
- **Neurobiological Correlates of Action Selection and Belief Updating**

Complete Taxonomy Tree

- Reducing Belief Deviation in Reinforcement Learning for Active Reasoning Survey Taxonomy
- LLM-Based Active Reasoning with Belief Tracking
  - Belief Deviation Control in Active Reasoning ★ (3 papers)
  - [0] $\mathbf{T^3}$: Reducing Belief Deviation in Reinforcement Learning for Active Reasoning (Anon et al., 2026) View paper
  - [1] : Reducing Belief Deviation in Reinforcement Learning for Active Reasoning (D Zou, 2025) View paper
  - Consistency-Based Self-Rewarding for LLM Reasoning (1 papers)
  - [2] Consistent Paths Lead to Truth: Self-Rewarding Reinforcement Learning for LLM Reasoning (Yao Qi, 2025) View paper
  - Token Efficiency and Response Conciseness in RL-Trained Reasoning (1 papers)
  - [3] Concise Reasoning via Reinforcement Learning (Fatemi, 2025) View paper
- Multiagent Epistemic Planning and Belief Coordination
  - Epistemic Planning in Multiagent Deep RL (1 papers)
  - [7] Consistent epistemic planning for multiagent deep reinforcement learning (Peiliang Wu, 2023) View paper
  - AI-Enabled Scenario Planning for Policy-Making (1 papers)
  - [8] AI4Policy: AI-Enabled Scenario Planning for Policy-Making in the Age of AI (J Kgomo, 2025) View paper
- Active Inference and Belief Dynamics
  - Single-Agent Active Inference for Autonomous Control (1 papers)
  - [6] Active Inference for an Intelligent Agent in Autonomous Reconnaissance Missions (Schubert, 2025) View paper
  - Multiagent Active Inference and Belief Sharing (2 papers)
  - [10] Federated inference and belief sharing (K. Friston, 2023) View paper
  - [11] Multiagent model of active inference (Nouri, n.d.) View paper
- Hybrid Learning Systems Combining RL and Symbolic Reasoning (1 papers)
  - [9] Combining Reinforcement Learning and Belief Revision-A Learning System for Active Vision. (T LÃ©opold, 2008) View paper
- Neurobiological Correlates of Action Selection and Belief Updating (1 papers)
  - [5] Variability in action selection relates to striatal dopamine 2/3 receptor availability in humans: a PET neuroimaging study using reinforcement learning and active â (RA Adams, 2020) View paper

Narrative

Core task: Reducing belief deviation in reinforcement learning for active reasoning. This field addresses how agents maintain coherent internal beliefs while actively reasoning and acting in complex environments. The taxonomy reveals several complementary perspectives: LLM-Based Active Reasoning with Belief Tracking explores how large language models can be guided to reason more reliably by monitoring and correcting belief inconsistencies; Multiagent Epistemic Planning and Belief Coordination examines how multiple agents

coordinate their knowledge and beliefs during collaborative tasks; Active Inference and Belief Dynamics draws on neuroscience-inspired frameworks where agents minimize prediction errors; Hybrid Learning Systems Combining RL and Symbolic Reasoning integrate neural and symbolic methods to ground beliefs in structured knowledge; and Neurobiological Correlates of Action Selection and Belief Updating connects computational models to brain mechanisms. Representative works such as Belief Deviation Reduction[1] and T3 Active Reasoning[4] illustrate how belief tracking can be operationalized within LLM-based reasoning pipelines, while Epistemic Planning Multiagent[7] and Multiagent Active Inference[11] highlight coordination challenges in multi-agent settings.

A particularly active line of work focuses on controlling belief drift during iterative reasoning steps, where agents must balance exploration with maintaining consistency. T3 Belief Deviation[0] sits squarely within this cluster, emphasizing mechanisms to detect and reduce deviations as reasoning unfolds—closely aligned with Belief Deviation Reduction[1] and T3 Active Reasoning[4], which similarly target coherence in LLM-driven inference. In contrast, Concise Reasoning[3] prioritizes efficiency and brevity over exhaustive belief tracking, suggesting a trade-off between computational cost and epistemic rigor. Meanwhile, branches like Active Inference Reconnaissance[6] and Striatal Dopamine Selection[5] offer biologically grounded perspectives on how belief updating might be implemented in neural circuits, raising open questions about whether computational models can benefit from these neurobiological insights. The original paper's focus on belief deviation control positions it as a methodological contribution to ensuring robust active reasoning in LLM-based agents.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. : Reducing Belief Deviation in Reinforcement Learning for Active Reasoning

Authors: D Zou, Y Chen, J Wang, H Yang, M Li, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ Our findings establish belief deviation as a central bottleneck and show that controlling it is a principled pathway toward building robust and generalizable active reasoning agents. â¦

⚠ Similarity Notice

This paper appears to be the same as the original paper based on the identical title and matching content fragments describing belief deviation control in reinforcement learning for active reasoning. The candidate paper's abstract excerpt directly matches the core contribution and findings of the original paper.

---

2. T3: Reducing Belief Deviation in Reinforcement Learning for Active Reasoning

Authors: Zou Deyu, Chen Yongqiang, Deyu Zou, Wang Jian-Xiang, Yongqiang Chen, et al. (16 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Active reasoning requires large language models (LLMs) to interact with external sources and strategically gather information to solve problems. Central to this process is belief tracking: maintaining a coherent understanding of the problem state and the missing information toward the solution. However, due to limited reasoning capabilities, LLM-based agents often suffer from belief deviation: they struggle to correctly model beliefs, lose track of problem states, and fall into uninformative or ...

⚠ Similarity Notice

This paper appears to be the same work as the original paper, likely a different version or submission. Both papers share an identical title (T3: Reducing Belief Deviation in Reinforcement Learning for Active Reasoning), nearly identical abstracts describing the same T3 method for detecting and truncating belief-deviated trajectories in RL training, and report the same experimental results (up to 30% gains with token reduction). The only minor difference is the reported token reduction percentage (34% vs 25%), which suggests these may be different submission versions of the same work.

Contributions Analysis

Overall novelty summary. The paper introduces T³, a method for detecting and truncating belief-trapped trajectories during reinforcement learning training of LLM-based active reasoning agents. It resides in the 'Belief Deviation Control in Active Reasoning' leaf, which contains only three papers total, including this work and two siblings. This represents a relatively sparse research direction within the broader taxonomy of eleven papers across multiple branches, suggesting the specific problem of belief deviation control in LLM active reasoning is an emerging rather than saturated area.

The taxonomy reveals neighboring work in consistency-based self-rewarding and token efficiency optimization, both under the same parent branch of 'LLM-Based Active Reasoning with Belief Tracking'. Sibling approaches address related but distinct challenges: one focuses on self-rewarding frameworks leveraging trajectory consistency without explicit belief deviation control, while another emphasizes reducing token usage rather than tracking epistemic coherence. The taxonomy also shows parallel branches in multiagent epistemic planning and active inference frameworks, which address belief dynamics in different computational paradigms (multiagent coordination and neuroscience-inspired free energy minimization respectively), highlighting that belief tracking spans multiple methodological traditions.

Among seventeen candidates examined across three contributions, none were found to clearly refute any aspect of the proposed work. The T³ truncation method examined four candidates with zero refutable matches; the theoretical characterization of belief-trap regions examined ten candidates, also with no refutations; and the T³ condition as a detection proxy examined three candidates without finding prior overlap. This limited search scope—seventeen papers rather than an exhaustive survey—suggests the analysis captures top semantic matches but may not cover all relevant prior work in trajectory optimization or credit assignment for sequential decision-making.

Given the sparse taxonomy leaf and absence of refutations among examined candidates, the work appears to address a relatively underexplored intersection of belief tracking and RL trajectory management for LLMs. However, the seventeen-paper search scope is modest, and the taxonomy's focus on belief-centric methods may underrepresent broader RL literature on trajectory truncation, early stopping, or credit assignment that could inform or overlap with this contribution.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: T3 method for truncating belief-trapped trajectories in RL

Description: The authors introduce T3, a training method that identifies when an agent enters a belief-trap region during reinforcement learning and truncates the trajectory at that point. By removing uninformative trajectory segments, T3 preserves credit assignment for informative actions and improves policy optimization.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Multistep Credit Assignment in Deep Reinforcement Learning

URL: View paper

Brief Assessment

Multistep Credit Assignment[12] focuses on compound returns (λ-returns, n-step returns) and eligibility traces for temporal credit assignment in deep RL with experience replay. It does not address belief deviation, partial observability (POMDP), or trajectory truncation based on detecting belief-trap regions—the core mechanisms of T3.

---

2. : Reducing Belief Deviation in Reinforcement Learning for Active Reasoning

URL: View paper

Brief Assessment

Belief Deviation Reduction[1] focuses on reducing belief deviation in active reasoning tasks through trajectory truncation. The original paper's T3 method addresses belief-trap regions in partially observable environments, which is a distinct technical contribution not challenged by this candidate.

---

3. Drift method: from stochastic networks to machine learning

URL: View paper

Brief Assessment

Drift Method[13] focuses on stability analysis in stochastic networks and online learning using Lyapunov drift techniques. It does not address trajectory truncation in reinforcement learning for active reasoning or belief-trap detection in LLM agents.

---

4. The truncated conjugate gradient (TCG), a non-iterative/fixed-cost strategy for computing polarization in molecular dynamics: Fast evaluation of analytical â□¦

URL: View paper

Brief Assessment

Truncated Conjugate Gradient[14] addresses computational chemistry problems (polarization in molecular dynamics), not reinforcement learning for active reasoning. The truncation mechanism serves entirely different purposes in distinct domains.

Contribution 2: Theoretical characterization of belief-trap regions and their impact on credit assignment

Description: The authors formalize the concept of belief-trap regions in partially observable Markov decision processes and prove that imperfect belief modeling causes agents to enter absorbing regions where progress stalls. They further demonstrate that these regions corrupt credit assignment by inverting gradient estimates for early exploratory actions.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Transformer-Based Multi-Agent Reinforcement Learning Method With Credit-Oriented Strategy Differentiation

URL: View paper

Brief Assessment

Credit Oriented Strategy[22] addresses credit assignment in multi-agent reinforcement learning through attention-based value decomposition, not belief-trap regions in partially observable single-agent settings.

---

2. Priority Over Quantity: A Self-Incentive Credit Assignment Scheme for Cooperative Multiagent Reinforcement Learning

URL: View paper

Brief Assessment

Priority Over Quantity[15] addresses credit assignment in cooperative multiagent RL with full observability, not belief-trap regions in partially observable single-agent settings. The candidate focuses on value factorization methods for multiagent coordination, which is a fundamentally different problem domain from the original paper's analysis of belief deviation in POMDPs.

---

3. LERO: LLM-driven Evolutionary framework with Hybrid Rewards and Enhanced Observation for Multi-Agent Reinforcement Learning

URL: View paper

Brief Assessment

LERO Evolutionary Framework[16] addresses credit assignment in multi-agent cooperative tasks through reward decomposition, not belief-trap regions in partially observable single-agent settings. The candidate focuses on multi-agent coordination challenges rather than belief modeling failures in active reasoning.

---

4. : Reducing Belief Deviation in Reinforcement Learning for Active Reasoning

URL: View paper

Brief Assessment

Belief Deviation Reduction[1] does not provide theoretical characterization of belief-trap regions in POMDPs or prove how imperfect belief modeling causes absorbing regions that corrupt credit assignment.

---

5. A Multiagent Cooperative Learning System With Evolution of Social Roles

URL: View paper

Brief Assessment

Social Roles Evolution[18] addresses multiagent reinforcement learning with role-based credit assignment in cooperative tasks, not belief-trap regions in partially observable single-agent active reasoning scenarios.

---

6. Transformers in Reinforcement Learning: A Survey

URL: View paper

Brief Assessment

Transformers RL Survey[17] provides a broad overview of transformers in RL and discusses credit assignment as a general challenge in classical RL algorithms. However, it does not present theoretical characterizations of belief-trap regions in POMDPs or prove how imperfect belief modeling corrupts credit assignment through gradient inversion, which are the specific novel contributions of the original paper.

7. MARL-CC: A Mathematical Framework forMulti-Agent Reinforcement Learning in ConnectedAutonomous Vehicles: Addressing Nonlinearity,Partial Observability, and Credit Assignment forOptimal Control

URL: View paper

Brief Assessment

MARL Connected Vehicles[20] focuses on multi-agent coordination in autonomous vehicle systems with differential geometric control and Shapley-value credit assignment, not on belief-trap regions in partially observable single-agent settings or their corruption of gradient estimates during reinforcement learning.

8. Terra Nova: A Comprehensive Challenge Environment for Intelligent Agents

URL: View paper

Brief Assessment

Terra Nova Environment[21] focuses on creating a comprehensive challenge environment inspired by Civilization V that integrates multiple RL challenges simultaneously. It does not address belief-trap regions, partial observability in the context of belief modeling, or credit assignment corruption in the theoretical manner described in the original paper.

9. Multi-Strategy Distillation Based on CTCE and CEDE

URL: View paper

Brief Assessment

Multi Strategy Distillation[23] addresses credit assignment in multi-agent systems through knowledge distillation between teacher-student models, not through theoretical characterization of belief-trap regions in partially observable single-agent environments.

10. Efficient Recurrent Off-Policy RL Requires a Context-Encoder-Specific Learning Rate

URL: View paper

Brief Assessment

Context Encoder Learning[19] addresses gradient instability in recurrent RL through learning rate adjustments for RNN context encoders. It does not formalize belief-trap regions or analyze how imperfect belief modeling corrupts credit assignment in POMDPs.

Contribution 3: T3 condition as a practical proxy for detecting belief-trap entry

Description: The authors propose a general truncation condition based on detecting stalled progress in the hypothesis space through observable proxy signals. This condition provides a practical implementation of the theoretical truncation principle without requiring direct access to unobservable belief states or thresholds.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. The development of scientific reasoning in knowledge-rich contexts.

URL: View paper

Brief Assessment

Scientific Reasoning Development[25] studies human cognitive development in scientific reasoning contexts, focusing on hypothesis space evolution in children's learning. This is fundamentally different from the ORIGINAL paper's computational framework for detecting stalled progress in LLM-based reinforcement learning agents through observable proxy signals.

2. Spectral policy optimization: Coloring your incorrect reasoning in grpo

URL: View paper

Brief Assessment

Spectral Policy Optimization[24] addresses reward signal issues in all-negative-sample groups in GRPO through AI feedback-based process supervision, not detecting stalled progress in hypothesis spaces through reasoning traces.

3. : Reducing Belief Deviation in Reinforcement Learning for Active Reasoning

URL: View paper

Brief Assessment

Belief Deviation Reduction[1] does not propose a general truncation condition based on detecting stalled progress in hypothesis space through observable proxy signals as described in the original paper.

Appendix: Text Similarity Detection

Textual similarity detection checked 16 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. T3: Reducing Belief Deviation in Reinforcement Learning for Active Reasoning

Detected in: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] $\mathbf{T^3}$: Reducing Belief Deviation in Reinforcement Learning for Active Reasoning View paper
- [1] : Reducing Belief Deviation in Reinforcement Learning for Active Reasoning View paper
- [2] Consistent Paths Lead to Truth: Self-Rewarding Reinforcement Learning for LLM Reasoning View paper
- [3] Concise Reasoning via Reinforcement Learning View paper
- [4] T3: Reducing Belief Deviation in Reinforcement Learning for Active Reasoning View paper
- [5] Variability in action selection relates to striatal dopamine 2/3 receptor availability in humans: a PET neuroimaging study using reinforcement learning and active â View paper
- [6] Active Inference for an Intelligent Agent in Autonomous Reconnaissance Missions View paper
- [7] Consistent epistemic planning for multiagent deep reinforcement learning View paper
- [8] AI4Policy: AI-Enabled Scenario Planning for Policy-Making in the Age of AI View paper

- [9] Combining Reinforcement Learning and Belief Revision-A Learning System for Active Vision. View paper
- [10] Federated inference and belief sharing View paper
- [11] Multiagent model of active inference View paper
- [12] Multistep Credit Assignment in Deep Reinforcement Learning View paper
- [13] Drift method: from stochastic networks to machine learning View paper
- [14] The truncated conjugate gradient (TCG), a non-iterative/fixed-cost strategy for computing polarization in molecular dynamics: Fast evaluation of analytical â View paper
- [15] Priority Over Quantity: A Self-Incentive Credit Assignment Scheme for Cooperative Multiagent Reinforcement Learning View paper
- [16] LERO: LLM-driven Evolutionary framework with Hybrid Rewards and Enhanced Observation for Multi-Agent Reinforcement Learning View paper
- [17] Transformers in Reinforcement Learning: A Survey View paper
- [18] A Multiagent Cooperative Learning System With Evolution of Social Roles View paper
- [19] Efficient Recurrent Off-Policy RL Requires a Context-Encoder-Specific Learning Rate View paper
- [20] MARL-CC: A Mathematical Framework forMulti-Agent Reinforcement Learning in ConnectedAutonomous Vehicles: Addressing Nonlinearity,Partial Observability, and Credit Assignment forOptimal Control View paper
- [21] Terra Nova: A Comprehensive Challenge Environment for Intelligent Agents View paper
- [22] Transformer-Based Multi-Agent Reinforcement Learning Method With Credit-Oriented Strategy Differentiation View paper
- [23] Multi-Strategy Distillation Based on CTCE and CEDE View paper
- [24] Spectral policy optimization: Coloring your incorrect reasoning in grpo View paper
- [25] The development of scientific reasoning in knowledge-rich contexts. View paper