Novelty Assessment Report

Paper: Large Reasoning Models Learn Better Alignment from Flawed Thinking

PDF URL: https://openreview.net/pdf?id=TSk3cIdIQK

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

Large reasoning models (LRMs) “think” by generating structured chain-of-thought (CoT) before producing a final answer, yet they still lack the ability to reason critically about safety alignment and are easily biased when a flawed premise is injected into their thought process. We propose **RECAP** (Robust Safety Alignment via Counter-Aligned Prefilling), a principled reinforcement learning (RL) method for post-training that explicitly teaches models to override flawed reasoning trajectories and reroute to safe and helpful responses. RECAP trains on a mixture of synthetically generated counter-aligned CoT prefills and standard prompts, requires no additional training cost or modifications beyond vanilla reinforcement learning from human feedback (RLHF), and substantially improves safety and jailbreak robustness, reduces overrefusal, and preserves core reasoning capability — all while maintaining inference token budget. Extensive analysis shows that RECAP-trained models engage in self-reflection more frequently and remain robust under adaptive attacks, preserving safety even after repeated attempts to override their reasoning.

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: **Improving Safety Alignment in Large Reasoning Models through Counter-Aligned Chain-of-Thought Prefilling**

A total of **1 papers** were analyzed and organized into a taxonomy with **2 categories**.

Taxonomy Overview

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

- **Counter-Aligned Reasoning Trajectory Methods**
- **Adversarial Chain-of-Thought Tuning Methods**

Complete Taxonomy Tree

- Improving Safety Alignment in Large Reasoning Models through Counter-Aligned Chain-of-Thought Prefilling Survey Taxonomy
- Counter-Aligned Reasoning Trajectory Methods
  - Reinforcement Learning-Based Counter-Alignment ★ (1 papers)
  - [0] Large Reasoning Models Learn Better Alignment from Flawed Thinking (Anon et al., 2026) View paper
- Adversarial Chain-of-Thought Tuning Methods
  - Snowball Effect Mitigation Techniques (1 papers)
  - [1] AdvChain: Adversarial Chain-of-Thought Tuning for Robust Safety Alignment of Large Reasoning Models (Zhu, 2025) View paper

Narrative

Core task: improving safety alignment in large reasoning models through counter-aligned chain-of-thought prefilling. This emerging field addresses a critical challenge in modern AI systems: ensuring that models capable of sophisticated reasoning do not produce harmful outputs even when prompted adversarially. The taxonomy reveals two primary branches that capture distinct methodological approaches. Counter-Aligned Reasoning Trajectory Methods focus on generating and leveraging reasoning paths that deliberately expose or counteract misaligned behavior, often through reinforcement learning or other trajectory-optimization techniques. Adversarial Chain-of-Thought Tuning Methods, by contrast, emphasize adversarial training regimes that directly manipulate intermediate reasoning steps to stress-test and harden model safety. Together, these branches reflect a shared recognition that safety alignment must extend beyond surface-level output filtering to the internal reasoning processes themselves.

Within Counter-Aligned Reasoning Trajectory Methods, a particularly active line of work explores reinforcement learning-based counter-alignment, where models learn to recognize and avoid flawed reasoning patterns through iterative feedback. Flawed Thinking[0] exemplifies this direction by using counter-aligned chain-of-thought prefilling to guide models away from unsafe trajectories during inference. This approach contrasts with adversarial tuning strategies such as AdvChain[1], which instead injects adversarial reasoning chains during training to preemptively expose vulnerabilities. The central trade-off across these branches involves balancing the computational cost of generating diverse counter-aligned trajectories against the robustness gains achieved. Flawed Thinking[0] sits squarely within the reinforcement learning-based counter-alignment cluster, emphasizing inference-time intervention rather than adversarial pre-training, and thus offers a complementary perspective to methods that rely on adversarial chain manipulation during the tuning phase.

Related Works in Same Category

No sibling papers and no sibling subtopics were found under the same parent taxonomy node; the paper appears structurally isolated in the taxonomy.

Contributions Analysis

Overall novelty summary. The paper proposes RECAP, a reinforcement learning method that trains models to override flawed reasoning trajectories using counter-aligned chain-of-thought prefills. According to the taxonomy, this work occupies the 'Reinforcement Learning-Based Counter-Alignment' leaf under 'Counter-Aligned Reasoning Trajectory Methods'. Notably, this leaf contains only the original paper itself—no sibling papers are present. This positioning suggests the paper addresses a relatively sparse research direction within the broader field of safety alignment for reasoning models, though the taxonomy includes only two papers total across both major branches.

The taxonomy reveals two primary methodological branches: Counter-Aligned Reasoning Trajectory Methods (where this paper resides) and Adversarial Chain-of-Thought Tuning Methods. The latter includes work on 'Snowball Effect Mitigation Techniques' that prevent progressive amplification of reasoning deviations. The taxonomy's scope notes clarify that counter-aligned prefilling methods (like RECAP) differ from adversarial training approaches by focusing on overriding flawed premises rather than preventing reasoning deviation amplification. This structural separation indicates the paper explores a distinct intervention point—teaching models to self-correct during reasoning rather than hardening them against adversarial inputs during training.

Among 30 candidate papers examined, the contribution-level analysis shows mixed novelty signals. The core RECAP method (Contribution 1) examined 10 candidates with zero refutable matches, suggesting limited direct prior work on RL-based counter-aligned prefilling within the search scope. However, the claim of simultaneous improvement across safety, helpfulness, and reasoning (Contribution 2) found one refutable candidate among 10 examined, indicating some overlap with existing multi-objective alignment work. The robustness claim under adaptive attacks (Contribution 3) showed no refutations across 10 candidates, though this reflects the limited search scale rather than exhaustive coverage.

Given the constrained literature search (30 candidates from semantic search), the paper appears to introduce a relatively novel training paradigm within its specific methodological niche. The absence of sibling papers in the taxonomy leaf and the low refutation rate for the core method suggest meaningful differentiation from prior work, though the single-paper taxonomy structure limits confidence in assessing field saturation. The analysis captures top-K semantic matches but does not guarantee comprehensive coverage of all relevant safety alignment or reasoning model literature.

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

Contribution 1: RECAP: Robust Safety Alignment via Counter-Aligned Prefilling

Description: RECAP is a reinforcement learning method that trains large reasoning models on a mixture of counter-aligned chain-of-thought prefills and standard prompts. By exposing models to deliberately flawed reasoning traces during training, RECAP teaches them to recover from misleading trajectories without requiring additional training cost or modifications beyond standard RLHF.

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. Self-rewarding correction for mathematical reasoning

URL: View paper

Brief Assessment

Self-Rewarding Correction[31] focuses on mathematical reasoning with self-correction and self-rewarding mechanisms for detecting errors in mathematical solutions. RECAP addresses safety alignment in reasoning models by training them to override flawed safety-related reasoning trajectories through counter-aligned prefilling.

---

2. MARS-SQL: A multi-agent reinforcement learning framework for Text-to-SQL

URL: View paper

Brief Assessment

MARS-SQL[30] focuses on Text-to-SQL translation using multi-agent RL with database interaction and self-correction for query generation. This is a completely different domain (SQL generation) and problem setting (structured query synthesis) compared to RECAP's safety alignment for reasoning models through counter-aligned prefilling.

---

3. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms

URL: View paper

Brief Assessment

Verifiable Rewards[27] focuses on reinforcement learning with verifiable rewards for mathematical and coding tasks, emphasizing correct reasoning through answer verification. RECAP addresses safety alignment by training models to override flawed reasoning trajectories using counter-aligned prefills, which is a distinct problem domain and methodology.

---

4. Reinforcement Learning for Reasoning in Large Language Models with One Training Example

URL: View paper

Brief Assessment

One Training Example[28] focuses on mathematical reasoning improvement through reinforcement learning with verifiable rewards using minimal training data. It does not address safety alignment, counter-aligned prefilling, or teaching models to override flawed safety reasoning trajectories, which are the core contributions of RECAP.

---

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

URL: View paper

Brief Assessment

Spectral Policy[26] focuses on improving GRPO by diversifying rewards in all-negative-sample groups for mathematical reasoning tasks, not on teaching models to override flawed safety-related reasoning trajectories through counter-aligned prefilling.

---

6. Training language models to self-correct via reinforcement learning

URL: View paper

Brief Assessment

Self-Correct Training[23] focuses on teaching models to self-correct mathematical and coding errors through multi-turn RL, not on safety alignment via counter-aligned prefilling. The candidate addresses intrinsic self-correction for reasoning tasks, while RECAP specifically targets safety robustness by training on deliberately flawed safety-oriented reasoning traces.

---

7. ReST-MCTS*: LLM Self-Training via Process Reward Guided Tree Search

URL: View paper

Brief Assessment

ReST-MCTS[25] focuses on improving reasoning quality through tree search and process rewards for mathematical/logical tasks, not on safety alignment or teaching models to override flawed safety reasoning trajectories.

---

8. Demystifying Long Chain-of-Thought Reasoning in LLMs

URL: View paper

Brief Assessment

Long Chain Reasoning[22] focuses on understanding the mechanics of long chain-of-thought emergence through RL training for mathematical reasoning tasks, not on safety alignment or teaching models to override flawed reasoning trajectories. The candidate investigates factors enabling long CoT generation (SFT necessity, compute scaling, reward shaping), while RECAP addresses safety brittleness by training on counter-aligned prefills to recover from misleading reasoning.

9. The Surprising Effectiveness of Negative Reinforcement in LLM Reasoning

URL: View paper

Brief Assessment

Negative Reinforcement[29] focuses on mathematical reasoning tasks and decomposes RL signals into positive/negative sample reinforcement, showing that penalizing incorrect responses improves pass@k performance. RECAP addresses safety alignment by training models to override counter-aligned CoT prefills (unsafe reasoning for harmful prompts, overly conservative reasoning for benign prompts) to achieve robust safety behavior. These are fundamentally different objectives and mechanisms.

10. SimpleTIR: End-to-End Reinforcement Learning for Multi-Turn Tool-Integrated Reasoning

URL: View paper

Brief Assessment

SimpleTIR[24] focuses on stabilizing multi-turn tool-integrated reasoning through filtering void turns in RL training for math tasks. It does not address safety alignment, counter-aligned prefilling, or teaching models to override flawed safety reasoning trajectories, which are the core novelties of RECAP.

Contribution 2: Simultaneous improvement of safety, helpfulness, and reasoning capability

Description: RECAP delivers substantial gains across multiple dimensions: improved safety on direct harmful and jailbreaking benchmarks, reduced overrefusal on benign queries, and enhanced mathematical reasoning performance. These improvements are achieved while maintaining similar inference-time token budgets and are supported by theoretical analysis demonstrating higher expected reward under both prefilled and non-prefilled evaluation.

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. Intentionreasoner: Facilitating adaptive llm safeguards through intent reasoning and selective query refinement

URL: View paper

Brief Assessment

IntentionReasoner[8] focuses on guard model mechanisms for input/output filtering with multi-level classification and query rewriting, rather than post-training RL methods that simultaneously improve safety, helpfulness, and reasoning within the base model itself. The technical approaches and objectives differ fundamentally.

2. Deliberative alignment: Reasoning enables safer language models

URL: View paper

Brief Assessment

Deliberative Alignment[2] focuses on teaching models to reason over explicit safety specifications through chain-of-thought, achieving safety improvements primarily through specification adherence. The original paper's contribution centers on using counter-aligned prefilling in RL training to simultaneously improve safety, reduce overrefusal, and preserve math reasoning - a distinct technical approach not demonstrated in the candidate.

3. SAFER: Advancing Safety Alignment via Efficient Ex-Ante Reasoning

URL: View paper

Brief Assessment

SAFER[9] focuses on ex-ante reasoning with rule verification for safety alignment, while RECAP addresses brittleness in reasoning models through counter-aligned prefilling during RL training. The technical approaches and problem formulations differ fundamentally.

4. Reasoning-to-defend: Safety-aware reasoning can defend large language models from jailbreaking

URL: View paper

Brief Assessment

Reasoning to Defend[7] focuses on safety-aware reasoning during generation with pivot tokens for self-evaluation, rather than the ORIGINAL paper's approach of training on counter-aligned prefills to override flawed reasoning trajectories. The candidate addresses jailbreak defense through reasoning distillation and contrastive pivot optimization, while the ORIGINAL achieves multi-dimensional improvements through RECAP's counter-aligned prefilling strategy during RL training.

5. ChatGPT for good? On opportunities and challenges of large language models for education

URL: View paper

Brief Assessment

ChatGPT Education[3] focuses on educational applications of large language models, discussing opportunities and challenges for teaching and learning. It does not address simultaneous optimization of safety, helpfulness, and reasoning in language model training.

6. ERPO: Advancing Safety Alignment via Ex-Ante Reasoning Preference Optimization

URL: View paper

Prior Art Analysis

ERPO[6] demonstrates that their method achieves simultaneous improvements across safety, helpfulness, and reasoning capability, similar to the claims made in the original paper. Both papers report gains in safety metrics, reduced overrefusal on benign queries, and maintained or improved performance on reasoning tasks. ERPO[6] explicitly states achieving improvements in safety while 'maintaining helpfulness and response efficiency' and shows results where safety, utility, and general performance are jointly enhanced through their two-stage training approach.

Evidence

Evidence 1 - **Rationale**: ERPO[6] demonstrates that their method maintains general performance while improving safety, showing simultaneous enhancement across multiple dimensions similar to the original paper's claims. - **Original**: recap simultaneously strengthens safety, helpfulness, and math reasoning capability - **Candidate**: safer does not degrade general performance. balancing safety and helpfulness is crucial. as shown in table 3, sft often compromise general ability, performing worse than the original chat model on most benchmarks. backtrack struggles on mt-bench, gsm8k, and xstest. by contrast, dpo, c2-syn, star-1, ...

Evidence 2 - **Rationale**: Both papers report quantitative improvements in reducing overrefusal while maintaining safety, demonstrating similar multi-dimensional optimization achievements. - **Original**: recap achieves on average +12.3% on direct harmful benchmarks, +21.0% on jailbreaking benchmarks, and +7.8% on the helpfulness score for overrefusal. additionally, it improves math reasoning by +0.9% - **Candidate**: safer achieves an 8.5% higher appropriate response rate than the chat model on xstest, a benchmark with benign queries containing subtle safety triggers. this suggests safer helps the model accurately assess intent and avoid both refusal and over-refusal.

7. Responsible AI in construction safety: Systematic evaluation of large language models and prompt engineering

URL: View paper

Brief Assessment

Construction Safety AI[5] focuses on evaluating LLM performance on construction safety certification exams across multiple knowledge areas. It does not address simultaneous improvement of safety alignment, helpfulness (overrefusal reduction), and reasoning capability through training methods like RECAP does.

8. Enhancing AI Trustworthiness Through Automated Reasoning: A Novel Method for Explaining Deep Learning and LLM Reasoning

URL: View paper

Brief Assessment

Automated Reasoning Trust[11] focuses on explaining deep learning and LLM reasoning through symbolic abductive reasoning and neural attention mechanisms for safety-critical applications. It does not address the simultaneous optimization of safety, helpfulness, and reasoning capability in language models through reinforcement learning methods.

9. Multi-expert prompting improves reliability, safety, and usefulness of large language models

URL: View paper

Brief Assessment

Multi-Expert Prompting[4] focuses on improving truthfulness, factuality, toxicity, and hurtfulness through multi-expert aggregation at inference time, not on training methods for reasoning models. The candidate does not address the specific challenge of training LRMs to recover from flawed reasoning trajectories or the counter-aligned prefilling approach that is central to the original paper's contribution.

10. Superficial safety alignment hypothesis

URL: View paper

Brief Assessment

Superficial Safety[10] focuses on identifying safety-critical neural components through pruning and freezing mechanisms during fine-tuning, rather than on RL-based training methods that simultaneously optimize multiple reward signals across safety, helpfulness, and reasoning domains as in the original paper.

Contribution 3: Persistent robustness under adaptive attacks via increased self-reflection

Description: RECAP-trained models demonstrate sustained safety even when subjected to adaptive attacks designed to bypass their self-reflection mechanisms, including full CoT hijacking and iterative prefill reset attacks. Analysis reveals that these models engage in self-reflection significantly more frequently than vanilla RLHF models, actively revising unsafe or mistaken reasoning during generation.

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. Breaking agents: Compromising autonomous llm agents through malfunction amplification

URL: View paper

Brief Assessment

Breaking Agents[15] focuses on attacking LLM agents through malfunction amplification (causing repetitive/irrelevant actions), not on improving robustness through self-reflection mechanisms. The candidate examines vulnerabilities in agent systems, while the original contribution addresses safety alignment through counter-aligned prefilling training.

2. LARGO: Latent Adversarial Reflection through Gradient Optimization for Jailbreaking LLMs

URL: View paper

Brief Assessment

LARGO[12] focuses on jailbreaking LLMs through latent adversarial optimization, not on training models to resist attacks via self-reflection. The candidate paper is an attack method, while the original contribution describes a defense mechanism that increases self-reflection during training.

3. Latent syntax weaving in large language model representations: A novel mechanism for self-referential consistency in neural architectures

URL: View paper

Brief Assessment

Latent Syntax Weaving[17] focuses on self-referential consistency mechanisms in neural architectures and syntax weaving, not on adaptive attack robustness or self-reflection in safety alignment contexts as described in the original paper's contribution.

4. Quantitative self-reflection protocols for self-replicating memory chains in large language models: A technical investigation

URL: View paper

Brief Assessment

Self-Reflection Protocols[19] focuses on quantitative self-reflection protocols and memory chains in LLMs, examining model resilience to controlled input changes. It does not address adaptive attacks designed to bypass self-reflection mechanisms in safety-aligned reasoning models, nor does it discuss RLHF training methods or jailbreak robustness.

5. When to Trust Context: Self-Reflective Debates for Context Reliability

URL: View paper

Brief Assessment

Self-Reflective Debates[16] focuses on resolving conflicts between parametric knowledge and contextual input through multi-agent debate mechanisms, not on robustness under adaptive attacks designed to bypass self-reflection in reasoning models. The paper does not address adaptive attacks like CoT hijacking or iterative prefill reset attacks.

6. AegisLLM: Scaling Agentic Systems for Self-Reflective Defense in LLM Security

URL: View paper

Brief Assessment

AegisLLM[14] focuses on test-time multi-agent defense systems for jailbreaking and unlearning, not on training-time methods that induce self-reflection during generation. The candidate does not address adaptive attacks that target self-reflection mechanisms during model reasoning.

7. Chain-of-scrutiny: Detecting backdoor attacks for large language models

URL: View paper

Brief Assessment

Chain of Scrutiny[13] focuses on detecting backdoor attacks in LLMs through consistency verification between reasoning steps and outputs, not on training models to maintain safety under adaptive attacks through self-reflection mechanisms as RECAP does.

8. ShadowCoT: Cognitive Hijacking for Stealthy Reasoning Backdoors in LLMs

URL: View paper

Brief Assessment

ShadowCoT[21] focuses on attacking reasoning mechanisms through backdoor injection and cognitive hijacking, not on defending against adaptive attacks through self-reflection. The candidate demonstrates vulnerabilities in reasoning chains rather than robustness improvements.

9. Stochastic constraint self-reflective syntax reconstruction in large language model internal representational spaces

URL: View paper

Brief Assessment

Stochastic Constraint[18] focuses on constraint mechanisms in syntax reconstruction within LLM representational spaces, not on adaptive attack robustness or self-reflection mechanisms in reasoning models. The minimal context provided shows no discussion of adversarial attacks, safety alignment, or self-reflective reasoning dynamics.

10. Multi-agent LLM debate unveils the premise left unsaid

URL: View paper

Brief Assessment

Multi-Agent Debate[20] focuses on implicit premise recovery in argument mining through dialogic reasoning between LLM agents. It does not address adaptive attacks, safety alignment, or self-reflection mechanisms in reasoning models under adversarial conditions.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Large Reasoning Models Learn Better Alignment from Flawed Thinking View paper
- [1] AdvChain: Adversarial Chain-of-Thought Tuning for Robust Safety Alignment of Large Reasoning Models View paper
- [2] Deliberative alignment: Reasoning enables safer language models View paper
- [3] ChatGPT for good? On opportunities and challenges of large language models for education View paper
- [4] Multi-expert prompting improves reliability, safety, and usefulness of large language models View paper
- [5] Responsible AI in construction safety: Systematic evaluation of large language models and prompt engineering View paper
- [6] ERPO: Advancing Safety Alignment via Ex-Ante Reasoning Preference Optimization View paper
- [7] Reasoning-to-defend: Safety-aware reasoning can defend large language models from jailbreaking View paper
- [8] Intentionreasoner: Facilitating adaptive llm safeguards through intent reasoning and selective query refinement View paper
- [9] SAFER: Advancing Safety Alignment via Efficient Ex-Ante Reasoning View paper
- [10] Superficial safety alignment hypothesis View paper
- [11] Enhancing AI Trustworthiness Through Automated Reasoning: A Novel Method for Explaining Deep Learning and LLM Reasoning View paper
- [12] LARGO: Latent Adversarial Reflection through Gradient Optimization for Jailbreaking LLMs View paper
- [13] Chain-of-scrutiny: Detecting backdoor attacks for large language models View paper
- [14] AegisLLM: Scaling Agentic Systems for Self-Reflective Defense in LLM Security View paper
- [15] Breaking agents: Compromising autonomous llm agents through malfunction amplification View paper
- [16] When to Trust Context: Self-Reflective Debates for Context Reliability View paper
- [17] Latent syntax weaving in large language model representations: A novel mechanism for self-referential consistency in neural architectures View paper
- [18] Stochastic constraint self-reflective syntax reconstruction in large language model internal representational spaces View paper
- [19] Quantitative self-reflection protocols for self-replicating memory chains in large language models: A technical investigation View paper
- [20] Multi-agent LLM debate unveils the premise left unsaid View paper
- [21] ShadowCoT: Cognitive Hijacking for Stealthy Reasoning Backdoors in LLMs View paper
- [22] Demystifying Long Chain-of-Thought Reasoning in LLMs View paper
- [23] Training language models to self-correct via reinforcement learning View paper
- [24] SimpleTIR: End-to-End Reinforcement Learning for Multi-Turn Tool-Integrated Reasoning View paper
- [25] ReST-MCTS*: LLM Self-Training via Process Reward Guided Tree Search View paper
- [26] Spectral policy optimization: Coloring your incorrect reasoning in grpo View paper
- [27] Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms View paper
- [28] Reinforcement Learning for Reasoning in Large Language Models with One Training Example View paper
- [29] The Surprising Effectiveness of Negative Reinforcement in LLM Reasoning View paper

- [30] MARS-SQL: A multi-agent reinforcement learning framework for Text-to-SQL View paper
- [31] Self-rewarding correction for mathematical reasoning View paper