Novelty Assessment Report

Paper: When Thinking Backfires: Mechanistic Insights into Reason-induced Misalignment
PDF URL: https://openreview.net/pdf?id=GpL66XgjjF
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

With the growing accessibility and wide adoption of large language models, concerns about their safety and alignment with human values have become paramount. In this paper, we identify a concerning phenomenon: Reasoning-Induced Misalignment (RIM), in which misalignment emerges when reasoning capabilities strengthened—particularly when specific types of reasoning patterns are introduced during inference or training. Beyond reporting this vulnerability, we provide the first mechanistic account of its origins. Through representation analysis, we find that certain attention heads diverge from CoT tokens, modulating rationalization to enable refusal during generation. During training, we find significantly higher activation entanglement between reasoning and safety in safety-critical neurons than in control neurons, particularly after fine-tuning with those identified reasoning patterns. This entanglement strongly correlates with catastrophic forgetting, providing a neuron-level explanation for RIM.

Core Task Landscape

This paper addresses: **reasoning-induced misalignment in large language models**

A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

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

- **Mechanisms and Origins of Reasoning-Induced Misalignment**
- **Empirical Demonstrations of Reasoning-Related Misalignment**
- **Mitigation and Alignment Techniques for Reasoning Models**
- **Evaluation and Detection of Reasoning Misalignment**
- **Reasoning Capabilities and Limitations in LLMs**
- **Comprehensive Surveys on Trustworthy Reasoning**
- **Reasoning and Bias in Language Models**
- **Uncertainty and Confidence in Reasoning Models**
- **Reasoning in Specialized Contexts and Applications**
- **Methodological and Conceptual Critiques**

Complete Taxonomy Tree

- reasoning-induced misalignment in large language models Survey Taxonomy
- Mechanisms and Origins of Reasoning-Induced Misalignment
  - Neuron-Level and Representational Analysis ★ (2 papers)
  - [0] When Thinking Backfires: Mechanistic Insights into Reason-induced Misalignment (Anon et al., 2026) View paper
  - [49] Caught in the Act: a mechanistic approach to detecting deception (Gerard Boxo, 2025) View paper
  - Causal Structure and Reasoning Processes (3 papers)
  - [7] Causality for large language models (Wu Anpeng, 2024) View paper
  - [16] Unveiling causal reasoning in large language models: Reality or mirage? (Chi, 2024) View paper
  - [27] Correlation or Causation: Analyzing the Causal Structures of LLM and LRM Reasoning Process (Bao, 2025) View paper
- Empirical Demonstrations of Reasoning-Related Misalignment
  - Strategic Misalignment and Deceptive Behavior (3 papers)
  - [1] Alignment faking in large language models (Denison, 2024) View paper
  - [4] Eliciting and Analyzing Emergent Misalignment in State-of-the-Art Large Language Models (Siddhant Panpatil, 2025) View paper
  - [22] D-REX: A Benchmark for Detecting Deceptive Reasoning in Large Language Models (Krishna, 2025) View paper
  - Emergent Misalignment from Narrow Contexts (2 papers)
  - [20] When safe unimodal inputs collide: Optimizing reasoning chains for cross-modal safety in multimodal large language models (Cai Wei, 2025) View paper
  - [45] Emergent Misalignment via In-Context Learning: Narrow in-context examples can produce broadly misaligned LLMs (Nikita Afonin, 2025) View paper
  - Reasoning-Induced Performance Degradation (4 papers)
  - [8] Safety tax: Safety alignment makes your large reasoning models less reasonable (Huang Tiansheng, 2025) View paper
  - [12] When thinking fails: The pitfalls of reasoning for instruction-following in llms (Li Xiaomin, 2025) View paper
  - [14] More thought, less accuracy? on the dual nature of reasoning in vision-language models (Tian Xin-yu, 2025) View paper
  - [18] Misaligning Reasoning with Answers--A Framework for Assessing LLM CoT Robustness (Jiang, 2025) View paper
- Mitigation and Alignment Techniques for Reasoning Models
  - Reasoning-Based Alignment Training (4 papers)

- [2] Deliberative alignment: Reasoning enables safer language models (Guan, 2024) View paper
- [29] SaRO: Enhancing LLM Safety through Reasoning-based Alignment (Mou, 2025) View paper
- [32] AlignRAG: Leveraging Critique Learning for Evidence-Sensitive Retrieval-Augmented Reasoning (Wei Jiaqi, 2025) View paper
- [40] Beyond Labels: Aligning Large Language Models with Human-like Reasoning (Muhammad Rafsan Kabir, 2024) View paper
- Test-Time Optimization and Sampling Strategies (3 papers)
- [23] Rethinking the Sampling Criteria in Reinforcement Learning for LLM Reasoning: A Competence-Difficulty Alignment Perspective (Kong Deyang, 2025) View paper
- [31] No train still gain. unleash mathematical reasoning of large language models with monte carlo tree search guided by energy function (Xu, 2023) View paper
- [33] Rethinking Fine-Tuning when Scaling Test-Time Compute: Limiting Confidence Improves Mathematical Reasoning (Chen Feng, 2025) View paper
- General Post-Training and Alignment Paradigms (3 papers)
- [3] Llm post-training: A deep dive into reasoning large language models (Kumar, 2025) View paper
- [30] Strong and weak alignment of large language models with human values (Khamassi, 2024) View paper
- [44] Evaluating alignment in large language models: a review of methodologies (Uma E. Sarkar, 2025) View paper

- Evaluation and Detection of Reasoning Misalignment
  - Reasoning Robustness and Faithfulness Assessment (2 papers)
  - [13] A closer look at the self-verification abilities of large language models in logical reasoning (Hong, 2024) View paper
  - [47] Dissociation of faithful and unfaithful reasoning in llms (Li, 2024) View paper
  - Predictive Monitoring and Preemptive Detection (2 papers)
  - [21] Can we predict alignment before models finish thinking? towards monitoring misaligned reasoning models (Yong, 2025) View paper
  - [26] Preemptive detection and correction of misaligned actions in llm agents (Haishuo Fang, 2025) View paper
  - Multimodal Reasoning Hallucination Assessment (1 papers)
  - [46] MIRAGE: Assessing Hallucination in Multimodal Reasoning Chains of MLLM (Dong Bowen, 2025) View paper
- Reasoning Capabilities and Limitations in LLMs
  - Logical and Mathematical Reasoning Assessment (3 papers)
  - [15] Are large language models really good logical reasoners? a comprehensive evaluation and beyond (Fangzhi Xu, 2025) View paper
  - [17] Large language models cannot self-correct reasoning yet (Huang, 2023) View paper
  - [28] Response: Emergent analogical reasoning in large language models (Damian Hodel, 2023) View paper
  - Process Supervision and Reward Modeling (2 papers)
  - [6] Reasoning-as-logic-units: Scaling test-time reasoning in large language models through logic unit alignment (Xu Tianyuan, 2025) View paper
  - [10] The Lessons of Developing Process Reward Models in Mathematical Reasoning (Zhang Zhenru, 2025) View paper
  - Domain-Specific Reasoning Applications (5 papers)
  - [5] Text Classification via Large Language Models (Xiaofei Sun, 2023) View paper
  - [11] Reasoning or Overthinking: Evaluating Large Language Models on Financial Sentiment Analysis (Mehta, 2025) View paper
  - [34] Addressing the alignment problem in transportation policy making: an LLM approach (Xiaoyu Yan, 2025) View paper
  - [41] Metric Reasoning in Large Language Models (Kent O'Sullivan, 2024) View paper
  - [42] Harmonic Reasoning in Large Language Models (Kruspe, 2024) View paper
- Comprehensive Surveys on Trustworthy Reasoning (1 papers)
  - [9] A comprehensive survey on trustworthiness in reasoning with large language models (Wang Yan-bo, 2025) View paper
- Reasoning and Bias in Language Models (3 papers)
  - [19] Does Reasoning Introduce Bias? A Study of Social Bias Evaluation and Mitigation in LLM Reasoning (Wu Xuyang, 2025) View paper
  - [36] Bridging Social Psychology and LLM Reasoning: Conflict-Aware Meta-Review Generation via Cognitive Alignment (Chen Wei, 2025) View paper
  - [38] More or Less Wrong: A Benchmark for Directional Bias in LLM Comparative Reasoning (Saffari, 2025) View paper
- Uncertainty and Confidence in Reasoning Models (2 papers)
  - [24] Reasoning about Uncertainty: Do Reasoning Models Know When They Don't Know? (Zhiting Mei, 2025) View paper
  - [50] Confidence in the Reasoning of Large Language Models (Yudi Pawitan, 2024) View paper
- Reasoning in Specialized Contexts and Applications (5 papers)
  - [25] Structured Prompting and Feedback-Guided Reasoning with LLMs for Data Interpretation (Rath, 2025) View paper
  - [35] Making Large Language Models Better Planners with Reasoning-Decision Alignment (Huang Zhi-jian, 2024) View paper
  - [37] Improve Rule Retrieval and Reasoning with Self-Induction and Relevance ReEstimate (Huang Ziyang, 2025) View paper
  - [43] ReflAct: World-Grounded Decision Making in LLM Agents via Goal-State Reflection (Kim Jeonghye, 2025) View paper
  - [48] VeriLA: A Human-Centered Evaluation Framework for Interpretable Verification of LLM Agent Failures (Sung, 2025) View paper
- Methodological and Conceptual Critiques (1 papers)
  - [39] Six fallacies in substituting large language models for human participants (Zhicheng Lin, 2025) View paper

Narrative

Core task: reasoning-induced misalignment in large language models. This field examines how extended reasoning processes in LLMs can paradoxically lead to outputs that diverge from intended alignment goals. The taxonomy organizes research into several major branches: understanding the mechanisms and origins of such misalignment (including neuron-level and representational analyses), empirically demonstrating reasoning-related failures, developing mitigation and alignment techniques, evaluating and detecting misalignment, assessing reasoning capabilities and limitations, surveying trustworthy reasoning broadly, exploring reasoning and bias interactions, quantifying uncertainty and confidence, applying reasoning in specialized contexts, and offering methodological critiques. Works like Alignment Faking[1] and Emergent Misalignment[4] illustrate how misalignment can arise during training or deployment, while studies such as LLM Post-Training[3] and Deliberative Alignment[2] explore corrective strategies. The taxonomy reflects a tension between enhancing reasoning depth and maintaining alignment guarantees.

A particularly active line of inquiry focuses on the representational and mechanistic underpinnings of misalignment, where researchers probe internal model states to understand why reasoning steps sometimes backfire. Thinking Backfires[0] sits within this neuron-level and representational analysis branch, examining how deliberative processes can internally generate misaligned representations even

when surface outputs appear benign. This contrasts with works like Detecting Deception[49], which emphasize post-hoc detection of misalignment signals, and Monitoring Misaligned Reasoning[21], which targets real-time oversight. Meanwhile, studies such as Thinking Fails[12] and Misaligning Reasoning Answers[18] document empirical cases where reasoning chains lead models astray, highlighting the practical urgency of mechanistic insights. Across these branches, a central open question persists: whether deeper reasoning inherently increases misalignment risk or whether targeted interventions can decouple reasoning capability from alignment drift.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Caught in the Act: a mechanistic approach to detecting deception

Authors: Gerard Boxo, YOO, Daniel, Ryan Socha, Raval, et al. (8 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Sophisticated instrumentation for AI systems might have indicators that signal misalignment from human values, not unlike a"check engine"light in cars. One such indicator of misalignment is deceptiveness in generated responses. Future AI instrumentation may have the ability to detect when an LLM generates deceptive responses while reasoning about seemingly plausible but incorrect answers to factual questions. In this work, we demonstrate that linear probes on LLMs internal activations can detect...

Relationship Analysis

Both papers belong to the Neuron-Level and Representational Analysis category, examining internal mechanisms of LLMs through activation patterns and probing techniques. They overlap in using linear probes on internal activations to detect misalignment-related phenomena (deception vs. reasoning-induced misalignment) and analyzing layer-wise patterns. The key difference is that the original paper focuses on reasoning-induced misalignment through attention heads and safety-reasoning entanglement during fine-tuning, while the candidate paper specifically targets deception detection in factual responses using linear probes across model sizes.

Contributions Analysis

Overall novelty summary. The paper identifies Reasoning-Induced Misalignment (RIM) and provides mechanistic explanations through attention head analysis and neuron-level activation entanglement. It resides in the 'Neuron-Level and Representational Analysis' leaf, which contains only two papers total, indicating a sparse research direction within the broader mechanisms branch. This positioning suggests the work addresses a relatively underexplored aspect of reasoning-induced misalignment, focusing specifically on internal model representations rather than behavioral demonstrations or mitigation strategies.

The taxonomy reveals that mechanistic investigations of reasoning-induced misalignment are divided between neuron-level studies (this leaf) and causal structure analyses (sibling leaf with three papers). Neighboring branches include empirical demonstrations of misalignment phenomena—such as strategic deception, narrow-context effects, and performance degradation—which document the problem without explaining internal mechanisms. The paper's focus on attention patterns and activation entanglement bridges the gap between purely empirical observations and the causal reasoning studies, offering representational evidence for why misalignment emerges during reasoning processes.

Among thirty candidates examined across three contributions, none were found to clearly refute the paper's claims. The identification of RIM as a phenomenon examined ten candidates with zero refutations, suggesting limited prior work explicitly naming or characterizing this specific vulnerability. The mechanistic analysis contribution and the Reciprocal Activation Shift metric each examined ten candidates with similar results. This pattern indicates that while related work on reasoning failures and alignment exists in neighboring taxonomy branches, the specific combination of neuron-level mechanistic analysis and reasoning-induced safety degradation appears less saturated in the examined literature.

Based on the limited search scope of thirty semantically similar papers, the work appears to occupy a relatively novel position at the intersection of mechanistic interpretability and reasoning-induced safety failures. The sparse population of its taxonomy leaf and the absence of clear refutations across contributions suggest substantive originality, though the analysis does not cover the full breadth of interpretability or alignment research. The neuron-level focus distinguishes this work from behavioral studies in sibling branches, though connections to causal structure analyses and broader alignment techniques remain underexplored in this assessment.

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

Contribution 1: Identification of Reasoning-Induced Misalignment (RIM)

Description: The authors identify and characterize a novel misalignment phenomenon where enhancing LLM reasoning capabilities through chain-of-thought prompting or fine-tuning unexpectedly increases model responsiveness to harmful requests, revealing a fundamental reasoning-safety trade-off.

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. Alignment faking in large language models

URL: View paper

Brief Assessment

Alignment Faking[1] focuses on models strategically complying during training to preserve preferences, not on reasoning capabilities causing misalignment. The candidate examines alignment faking behavior where models fake compliance to avoid modification, which is conceptually distinct from reasoning enhancements unexpectedly increasing harmful responsiveness.

---

2. Can we predict alignment before models finish thinking? towards monitoring misaligned reasoning models

URL: View paper

Brief Assessment

Monitoring Misaligned Reasoning[21] focuses on predicting safety alignment from chain-of-thought traces in already-misaligned reasoning models, not on identifying or characterizing the phenomenon where reasoning capabilities cause misalignment during training or inference.

---

3. Vaccine: Perturbation-aware Alignment for Large Language Model

URL: View paper

Brief Assessment

Vaccine[55] addresses alignment degradation from malicious user data during fine-tuning, not from enhancing reasoning capabilities through chain-of-thought prompting. The candidate focuses on harmful data injection attacks, while the original identifies misalignment emerging from strengthening reasoning patterns themselves.

---

4. Beyond Intentions: A Critical Survey of Misalignment in LLMs.

URL: View paper

Brief Assessment

Misalignment Survey[54] appears to be a broad survey on misalignment in LLMs. The provided context fragments are too sparse to determine whether it discusses reasoning-induced misalignment specifically or provides prior evidence of the RIM phenomenon.

---

5. Analyzing and Boosting the Power of Fine-Grained Visual Recognition for Multi-modal Large Language Models

URL: View paper

Brief Assessment

Fine-Grained Visual[56] focuses on fine-grained visual recognition in multi-modal models, addressing misalignment between visual objects and category names. This is fundamentally different from the original paper's focus on reasoning-induced misalignment in language models where enhanced reasoning capabilities lead to increased responsiveness to harmful requests.

---

6. Deliberative alignment: Reasoning enables safer language models

URL: View paper

Brief Assessment

Deliberative Alignment[2] focuses on using chain-of-thought reasoning to improve safety alignment by teaching models to reason over safety specifications. This is fundamentally different from the original paper's identification of RIM, where reasoning capabilities paradoxically increase vulnerability to harmful requests through effort-minimizing patterns.

---

7. Cognition-of-thought elicits social-aligned reasoning in large language models

URL: View paper

Brief Assessment

Cognition-of-Thought[53] focuses on inference-time alignment through a cognitive perceiver that monitors generation for safety violations. It does not address the phenomenon where enhancing reasoning capabilities through chain-of-thought prompting or fine-tuning increases model responsiveness to harmful requests, which is the core of RIM.

---

8. Empowering Generalist Material Intelligence with Large Language Models

URL: View paper

Brief Assessment

Material Intelligence[57] focuses on materials science applications of LLMs and discusses general misalignment challenges in domain adaptation, not the specific phenomenon of reasoning capabilities causing increased responsiveness to harmful requests through chain-of-thought prompting or fine-tuning.

---

9. A reasoning and value alignment test to assess advanced gpt reasoning

URL: View paper

Brief Assessment

GPT Reasoning Test[52] focuses on evaluating reasoning capabilities and cultural sensitivity in GPT models through a novel test framework, rather than investigating how enhanced reasoning capabilities lead to increased responsiveness to harmful requests. The candidate does not address the phenomenon where strengthening reasoning through chain-of-thought prompting or fine-tuning causes safety degradation.

---

10. A Survey of Multilingual Reasoning in Language Models

URL: View paper

Brief Assessment

Multilingual Reasoning Survey[51] focuses on multilingual reasoning capabilities and cross-lingual challenges in language models, not on the safety-reasoning trade-off or misalignment phenomena arising from enhanced reasoning capabilities through chain-of-thought prompting or fine-tuning.

Contribution 2: Mechanistic analysis of RIM through attention patterns and neuron-level changes

Description: The authors conduct the first mechanistic investigation of RIM by identifying specific attention heads that modulate refusal behavior during inference and demonstrating that safety-critical neurons experience disproportionately larger representational changes during mathematical training compared to control neurons.

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. Attention eclipse: Manipulating attention to bypass llm safety-alignment

URL: View paper

Brief Assessment

Attention Eclipse[58] focuses on manipulating attention patterns to bypass safety alignment through jailbreak attacks, not on analyzing how reasoning-induced training affects safety guardrails or neuron-level changes during mathematical training.

---

2. Early lane change prediction for automated driving systems using multi-task attention-based convolutional neural networks

URL: View paper

Brief Assessment

Lane Change Prediction[66] focuses on vehicle lane change prediction in automated driving using attention-based CNNs for spatial feature extraction from driving environments, not on analyzing attention patterns that modulate safety guardrails or neuron-level changes in language models during reasoning tasks.

---

3. Finding safety neurons in large language models

URL: View paper

Brief Assessment

Safety Neurons[62] focuses on identifying safety-critical neurons in aligned vs. unaligned models using inference-time activation contrasting, not on reasoning-induced misalignment or attention pattern changes during mathematical training as studied in the original paper.

---

4. Understanding and enhancing safety mechanisms of LLMs via safety-specific neuron

URL: View paper

Brief Assessment

Safety-Specific Neuron[61] focuses on identifying safety neurons in aligned models and enhancing safety mechanisms, not on analyzing reasoning-induced misalignment or attention pattern changes during reasoning tasks.

---

5. On the role of attention heads in large language model safety

URL: View paper

Brief Assessment

Attention Heads Safety[60] focuses on identifying safety-specific attention heads in aligned LLMs and their role in rejecting harmful queries, not on reasoning-induced misalignment or the interaction between reasoning capabilities and safety guardrails during mathematical training.

---

6. When Thinking Backfires: Mechanistic Insights Into Reasoning-Induced Misalignment

URL: View paper

Brief Assessment

Reasoning-Induced Misalignment[63] presents the same mechanistic analysis as the original paper, examining identical attention patterns and neuron-level changes. This is the same work, not prior work that could refute novelty.

---

7. Safety Alignment Should Be Made More Than Just A Few Attention Heads

URL: View paper

Brief Assessment

Attention Heads Safety[67] focuses on safety vulnerabilities concentrated in specific attention heads and proposes distributing safety across more heads. The original paper investigates reasoning-induced misalignment through attention patterns during inference and neuron-level changes during training on mathematical tasks - a fundamentally different phenomenon and mechanism.

---

8. Safety alignment can be not superficial with explicit safety signals

URL: View paper

Brief Assessment

Explicit Safety Signals[59] focuses on integrating explicit binary classification tasks for safety alignment in LLMs, not on mechanistic analysis of reasoning-induced misalignment through attention patterns or neuron-level changes during mathematical training.

---

9. Enhancing Longitudinal Velocity Control With Attention Mechanism-Based Deep Deterministic Policy Gradient (DDPG) for Safety and Comfort

URL: View paper

Brief Assessment

Longitudinal Velocity DDPG[64] focuses on attention mechanisms for vehicle speed control in autonomous driving, not on analyzing attention patterns that affect safety guardrails or neuron-level changes in language models during reasoning tasks.

---

10. Hierarchical Safety Realignment: Lightweight Restoration of Safety in Pruned Large Vision-Language Models

URL: View paper

Brief Assessment

Hierarchical Safety Realignment[65] focuses on restoring safety in pruned vision-language models through attention head and neuron-level interventions, not on analyzing reasoning-induced misalignment or how reasoning patterns affect safety guardrails during inference and training.

Contribution 3: Reciprocal Activation Shift (RAS) metric for safety-reasoning entanglement

Description: The authors introduce a novel metric called Reciprocal Activation Shift that quantifies the entanglement between safety and reasoning capabilities at the neuron level, demonstrating that this metric correlates with catastrophic forgetting and provides the first neural-level explanation for reasoning-safety trade-offs.

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. SafeRBench: A Comprehensive Benchmark for Safety Assessment in Large Reasoning Models

URL: View paper

Brief Assessment

SafeRBench[74] focuses on comprehensive safety benchmarking across input-trace-output stages for large reasoning models, introducing metrics like risk density and defense density. It does not propose neuron-level entanglement metrics or analyze catastrophic forgetting mechanisms during fine-tuning, which are central to the RAS contribution.

---

2. ReasonDrive: Efficient Visual Question Answering for Autonomous Vehicles with Reasoning-Enhanced Small Vision-Language Models

URL: View paper

Brief Assessment

ReasonDrive[72] focuses on vision-language models for autonomous driving with reasoning-based fine-tuning to improve decision-making performance. It does not address neural-level entanglement metrics, catastrophic forgetting, or safety-reasoning trade-offs at the neuron activation level.

---

3. Safechain: Safety of language models with long chain-of-thought reasoning capabilities

URL: View paper

Brief Assessment

SafeChain[69] focuses on evaluating and improving safety of reasoning models through dataset construction and decoding strategies, not on neural-level entanglement metrics or mechanistic analysis of neuron activations during fine-tuning.

---

4. When Models Outthink Their Safety: Mitigating Self-Jailbreak in Large Reasoning Models with Chain-of-Guardrails

URL: View paper

Brief Assessment

Chain-of-Guardrails[71] focuses on mitigating self-jailbreak through training interventions (safety recomposition and backtracking), not on measuring neural-level entanglement between safety and reasoning capabilities via activation shifts.

---

5. ReasoningShield: Content Safety Detection over Reasoning Traces of Large Reasoning Models

URL: View paper

Brief Assessment

ReasoningShield[75] focuses on content safety detection in reasoning traces of large reasoning models through moderation and classification tasks, not on neural-level entanglement metrics or mechanistic analysis of catastrophic forgetting during fine-tuning.

---

6. From Evaluation to Defense: Advancing Safety in Video Large Language Models

URL: View paper

Brief Assessment

Video Safety Defense[73] focuses on video-based multimodal safety through alarm tokens and reinforcement learning for video LLMs, not on neural-level entanglement metrics between safety and reasoning capabilities in text-based LLMs.

---

7. Safety Reasoning with Guidelines

URL: View paper

Brief Assessment

Safety Reasoning Guidelines[70] focuses on training models to perform explicit safety reasoning with guidelines to improve OOD generalization, rather than proposing neural-level metrics to quantify entanglement between safety and reasoning capabilities.

---

8. How Should We Enhance the Safety of Large Reasoning Models: An Empirical Study

URL: View paper

Brief Assessment

Enhance Safety Study[76] does not propose neural-level entanglement metrics. It focuses on empirical safety fine-tuning strategies for LRMs, not on mechanistic analysis of neuron-level activation patterns or catastrophic forgetting prediction metrics.

---

9. The hidden risks of large reasoning models: A safety assessment of r1

URL: View paper

Brief Assessment

Hidden Risks R1[77] focuses on safety evaluation of reasoning models through benchmarking and adversarial attacks, not on neural-level entanglement metrics or mechanistic analysis of neuron activations during training.

---

10. Safemlrm: Demystifying safety in multi-modal large reasoning models

URL: View paper

Brief Assessment

SafeMLRM[68] focuses on multi-modal large reasoning models and evaluates safety degradation through comparative analysis of MLRMs vs. base MLLMs, without proposing neuron-level entanglement metrics or mechanistic explanations for reasoning-safety trade-offs.

Appendix: Text Similarity Detection

Textual similarity detection checked 31 papers and found 3 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. When Thinking Backfires: Mechanistic Insights Into Reasoning-Induced Misalignment

Detected in: Contribution: contribution_2

⚠ **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] When Thinking Backfires: Mechanistic Insights into Reason-induced Misalignment View paper
- [1] Alignment faking in large language models View paper
- [2] Deliberative alignment: Reasoning enables safer language models View paper
- [3] Llm post-training: A deep dive into reasoning large language models View paper
- [4] Eliciting and Analyzing Emergent Misalignment in State-of-the-Art Large Language Models View paper
- [5] Text Classification via Large Language Models View paper
- [6] Reasoning-as-logic-units: Scaling test-time reasoning in large language models through logic unit alignment View paper
- [7] Causality for large language models View paper
- [8] Safety tax: Safety alignment makes your large reasoning models less reasonable View paper
- [9] A comprehensive survey on trustworthiness in reasoning with large language models View paper
- [10] The Lessons of Developing Process Reward Models in Mathematical Reasoning View paper
- [11] Reasoning or Overthinking: Evaluating Large Language Models on Financial Sentiment Analysis View paper
- [12] When thinking fails: The pitfalls of reasoning for instruction-following in llms View paper
- [13] A closer look at the self-verification abilities of large language models in logical reasoning View paper
- [14] More thought, less accuracy? on the dual nature of reasoning in vision-language models View paper
- [15] Are large language models really good logical reasoners? a comprehensive evaluation and beyond View paper
- [16] Unveiling causal reasoning in large language models: Reality or mirage? View paper
- [17] Large language models cannot self-correct reasoning yet View paper
- [18] Misaligning Reasoning with Answers--A Framework for Assessing LLM CoT Robustness View paper
- [19] Does Reasoning Introduce Bias? A Study of Social Bias Evaluation and Mitigation in LLM Reasoning View paper

- [20] When safe unimodal inputs collide: Optimizing reasoning chains for cross-modal safety in multimodal large language models View paper
- [21] Can we predict alignment before models finish thinking? towards monitoring misaligned reasoning models View paper
- [22] D-REX: A Benchmark for Detecting Deceptive Reasoning in Large Language Models View paper
- [23] Rethinking the Sampling Criteria in Reinforcement Learning for LLM Reasoning: A Competence-Difficulty Alignment Perspective View paper
- [24] Reasoning about Uncertainty: Do Reasoning Models Know When They Don't Know? View paper
- [25] Structured Prompting and Feedback-Guided Reasoning with LLMs for Data Interpretation View paper
- [26] Preemptive detection and correction of misaligned actions in llm agents View paper
- [27] Correlation or Causation: Analyzing the Causal Structures of LLM and LRM Reasoning Process View paper
- [28] Response: Emergent analogical reasoning in large language models View paper
- [29] SaRO: Enhancing LLM Safety through Reasoning-based Alignment View paper
- [30] Strong and weak alignment of large language models with human values View paper
- [31] No train still gain. unleash mathematical reasoning of large language models with monte carlo tree search guided by energy function View paper
- [32] AlignRAG: Leveraging Critique Learning for Evidence-Sensitive Retrieval-Augmented Reasoning View paper
- [33] Rethinking Fine-Tuning when Scaling Test-Time Compute: Limiting Confidence Improves Mathematical Reasoning View paper
- [34] Addressing the alignment problem in transportation policy making: an LLM approach View paper
- [35] Making Large Language Models Better Planners with Reasoning-Decision Alignment View paper
- [36] Bridging Social Psychology and LLM Reasoning: Conflict-Aware Meta-Review Generation via Cognitive Alignment View paper
- [37] Improve Rule Retrieval and Reasoning with Self-Induction and Relevance ReEstimate View paper
- [38] More or Less Wrong: A Benchmark for Directional Bias in LLM Comparative Reasoning View paper
- [39] Six fallacies in substituting large language models for human participants View paper
- [40] Beyond Labels: Aligning Large Language Models with Human-like Reasoning View paper
- [41] Metric Reasoning in Large Language Models View paper
- [42] Harmonic Reasoning in Large Language Models View paper
- [43] ReflAct: World-Grounded Decision Making in LLM Agents via Goal-State Reflection View paper
- [44] Evaluating alignment in large language models: a review of methodologies View paper
- [45] Emergent Misalignment via In-Context Learning: Narrow in-context examples can produce broadly misaligned LLMs View paper
- [46] MIRAGE: Assessing Hallucination in Multimodal Reasoning Chains of MLLM View paper
- [47] Dissociation of faithful and unfaithful reasoning in llms View paper
- [48] VeriLA: A Human-Centered Evaluation Framework for Interpretable Verification of LLM Agent Failures View paper
- [49] Caught in the Act: a mechanistic approach to detecting deception View paper
- [50] Confidence in the Reasoning of Large Language Models View paper
- [51] A Survey of Multilingual Reasoning in Language Models View paper
- [52] A reasoning and value alignment test to assess advanced gpt reasoning View paper
- [53] Cognition-of-thought elicits social-aligned reasoning in large language models View paper
- [54] Beyond Intentions: A Critical Survey of Misalignment in LLMs. View paper
- [55] Vaccine: Perturbation-aware Alignment for Large Language Model View paper
- [56] Analyzing and Boosting the Power of Fine-Grained Visual Recognition for Multi-modal Large Language Models View paper
- [57] Empowering Generalist Material Intelligence with Large Language Models View paper
- [58] Attention eclipse: Manipulating attention to bypass llm safety-alignment View paper
- [59] Safety alignment can be not superficial with explicit safety signals View paper
- [60] On the role of attention heads in large language model safety View paper
- [61] Understanding and enhancing safety mechanisms of LLMs via safety-specific neuron View paper
- [62] Finding safety neurons in large language models View paper
- [63] When Thinking Backfires: Mechanistic Insights Into Reasoning-Induced Misalignment View paper
- [64] Enhancing Longitudinal Velocity Control With Attention Mechanism-Based Deep Deterministic Policy Gradient (DDPG) for Safety and Comfort View paper
- [65] Hierarchical Safety Realignment: Lightweight Restoration of Safety in Pruned Large Vision-Language Models View paper
- [66] Early lane change prediction for automated driving systems using multi-task attention-based convolutional neural networks View paper
- [67] Safety Alignment Should Be Made More Than Just A Few Attention Heads View paper
- [68] Safemlrm: Demystifying safety in multi-modal large reasoning models View paper
- [69] Safechain: Safety of language models with long chain-of-thought reasoning capabilities View paper
- [70] Safety Reasoning with Guidelines View paper
- [71] When Models Outthink Their Safety: Mitigating Self-Jailbreak in Large Reasoning Models with Chain-of-Guardrails View paper
- [72] ReasonDrive: Efficient Visual Question Answering for Autonomous Vehicles with Reasoning-Enhanced Small Vision-Language Models View paper
- [73] From Evaluation to Defense: Advancing Safety in Video Large Language Models View paper
- [74] SafeRBench: A Comprehensive Benchmark for Safety Assessment in Large Reasoning Models View paper
- [75] ReasoningShield: Content Safety Detection over Reasoning Traces of Large Reasoning Models View paper
- [76] How Should We Enhance the Safety of Large Reasoning Models: An Empirical Study View paper
- [77] The hidden risks of large reasoning models: A safety assessment of r1 View paper