Novelty Assessment Report

Paper: FaithCoT-Bench: Benchmarking Instance-Level Faithfulness of Chain-of-Thought Reasoning

PDF URL: https://openreview.net/pdf?id=lN3yKqqzF1

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

Large language models (LLMs) increasingly rely on Chain-of-Thought (CoT) prompting to improve problem-solving and provide seemingly transparent explanations. However, growing evidence shows that CoT often fail to faithfully represent the underlying reasoning process, raising concerns about their reliability in high-risk applications. Although prior studies have focused on mechanism-level analyses showing that CoTs can be unfaithful, they leave open the practical challenge of deciding whether a specific trajectory is faithful to the internal reasoning of the model. To address this gap, we introduce FaithCoT-Bench, a unified benchmark for instance-level CoT unfaithfulness detection. Our framework establishes a rigorous task formulation that formulates unfaithfulness detection as a discriminative decision problem, and provides FINE-CoT (Faithfulness instance evaluation for Chain-of-Thought), an expert-annotated collection of over 1,000 trajectories generated by four representative LLMs across four domains, including more than 300 unfaithful instances with fine-grained causes and step-level evidence. We further conduct a systematic evaluation of eleven representative detection methods spanning counterfactual, logit-based, and LLM-as-judge paradigms, deriving empirical insights that clarify the strengths and weaknesses of existing approaches and reveal the increased challenges of detection in knowledge-intensive domains and with more advanced models. To the best of our knowledge, FaithCoT-Bench establishes the first comprehensive benchmark for instance-level CoT faithfulness, setting a solid basis for future research toward more interpretable and trustworthy reasoning in LLMs.

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: **Instance-Level Faithfulness Detection of Chain-of-Thought Reasoning**

A total of **50 papers** were analyzed and organized into a taxonomy with **18 categories**.

Taxonomy Overview

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

- **Faithfulness Measurement and Detection Methods**
- **Verification and Correctness Assessment**
- **Reasoning Improvement and Training Methods**
- **Analysis and Understanding of CoT Reasoning**
- **Specialized Reasoning Tasks and Benchmarks**

Complete Taxonomy Tree

- Instance-Level Faithfulness Detection of Chain-of-Thought Reasoning Survey Taxonomy
- Faithfulness Measurement and Detection Methods
  - Counterfactual and Intervention-Based Detection (3 papers)
  - [3] Measuring faithfulness of chains of thought by unlearning reasoning steps (Martin Tutek, 2025) View paper
  - [4] Frit: Using causal importance to improve chain-of-thought faithfulness (Swaroop, 2025) View paper
  - [23] Measuring chain of thought faithfulness by unlearning reasoning steps (Martin Tutek, 2025) View paper
  - Probabilistic and Statistical Guarantees (1 papers)
  - [9] Probabilistic soundness guarantees in llm reasoning chains (You, 2025) View paper
  - Comprehensive Benchmarking and Evaluation Frameworks ★ (3 papers)
  - [0] FaithCoT-Bench: Benchmarking Instance-Level Faithfulness of Chain-of-Thought Reasoning (Anon et al., 2026) View paper
  - [15] On the difficulty of faithful chain-of-thought reasoning in large language models (SH Tanneru, 2024) View paper
  - [21] Towards better chain-of-thought: A reflection on effectiveness and faithfulness (Jiachun Li, 2025) View paper
  - Mechanistic and Interpretability Analysis (2 papers)
  - [14] Towards a Mechanistic Interpretation of Multi-Step Reasoning Capabilities of Language Models (Bosselut, 2023) View paper
  - [42] A mechanistic analysis of a transformer trained on a symbolic multi-step reasoning task (Bartelt, 2024) View paper
- Verification and Correctness Assessment
  - Formal and Symbolic Verification (4 papers)
  - [1] Faithful logical reasoning via symbolic chain-of-thought (Fei Hao, 2024) View paper
  - [2] Deductive Verification of Chain-of-Thought Reasoning (Ling Zhan, 2023) View paper
  - [12] Step-Wise Formal Verification for LLM-Based Mathematical Problem Solving (Zhou Kuo, 2025) View paper
  - [33] HoarePrompt: Structural Reasoning About Program Correctness in Natural Language (DAI Yihan, 2025) View paper
  - Process Reward Models and Step-Level Supervision (4 papers)
  - [5] The Lessons of Developing Process Reward Models in Mathematical Reasoning (Zhang Zhenru, 2025) View paper
  - [6] Improve Mathematical Reasoning in Language Models by Automated Process Supervision (Luo, 2024) View paper
  - [11] Hard2Verify: A Step-Level Verification Benchmark for Open-Ended Frontier Math (Pandit, 2025) View paper
  - [18] Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning (Setlur, 2024) View paper
  - Self-Verification and Zero-Shot Checking (3 papers)

  - [7] Zero-Shot Verification-guided Chain of Thoughts (Chowdhury, 2025) View paper
  - [22] VerifiAgent: a Unified Verification Agent in Language Model Reasoning (Jiuzhou Han, 2025) View paper
  - [30] Selfcheck: Using llms to zero-shot check their own step-by-step reasoning (Miao Ning, 2023) View paper
  - Domain-Specific Verification Applications (3 papers)
  - [16] Stepwise Verification and Remediation of Student Reasoning Errors with Large Language Model Tutors (Daheim, 2024) View paper
  - [17] LegalReasoner: Step-wised Verification-Correction for Legal Judgment Reasoning (Weijie Shi, 2025) View paper
  - [31] RelCheck: Improving Relation Extraction with Ontology-Guided and LLM-Based Validation (Mounir Ourekouch, 2025) View paper
- Reasoning Improvement and Training Methods
  - Reinforcement Learning for Reasoning (4 papers)
  - [13] RL on Incorrect Synthetic Data Scales the Efficiency of LLM Math Reasoning by Eight-Fold (Setlur, 2024) View paper
  - [32] R3-RAG: Learning Step-by-Step Reasoning and Retrieval for LLMs via Reinforcement Learning (Li Yuan, 2025) View paper
  - [39] RL Tango: Reinforcing Generator and Verifier Together for Language Reasoning (Zha Kaiwen, 2025) View paper
  - [47] Improving LLM Reasoning through Scaling Inference Computation with Collaborative Verification (Liang, 2024) View paper
  - Prompting and Decomposition Strategies (4 papers)
  - [10] Question decomposition improves the faithfulness of model-generated reasoning (Ansh Radhakrishnan, 2023) View paper
  - [41] Rex-Thinker: Grounded Object Referring via Chain-of-Thought Reasoning (Jiang Qing, 2025) View paper
  - [43] SGEU: enhancing LLM reasoning via backward exemplar generation and verification: Z. Wang et al. (Z Wang, 2025) View paper
  - [49] LaRS: Latent Reasoning Skills for Chain-of-Thought Reasoning (Qi Yanjun, 2024) View paper
  - Model Specialization and Fine-Tuning (3 papers)
  - [20] Specializing smaller language models towards multi-step reasoning (Fu Yao, 2023) View paper
  - [44] On the Impact of Fine-Tuning on Chain-of-Thought Reasoning (Elita Lobo, 2024) View paper
  - [45] Faithful Chain-of-Thought Reasoning (Qing Lyu, 2023) View paper
  - Reasoning Compression and Efficiency (2 papers)
  - [8] ConCISE: Confidence-guided Compression in Step-by-step Efficient Reasoning (Deng YongHeng, 2025) View paper
  - [27] R1-compress: Long chain-of-thought compression via chunk compression and search (Wang Yibo, 2025) View paper
- Analysis and Understanding of CoT Reasoning
  - Effectiveness and Performance Analysis (2 papers)
  - [24] Multi-step problem solving through a verifier: An empirical analysis on model-induced process supervision (Zihan Wang, 2024) View paper
  - [37] Token Signature: Predicting Chain-of-Thought Gains with Token Decoding Feature in Large Language Models (Liu Peijie, 2025) View paper
  - Faithfulness and Bias Characterization (3 papers)
  - [28] Chain-of-thought is not explainability (F Barez, 2025) View paper
  - [34] A Closer Look at Bias and Chain-of-Thought Faithfulness of Large (Vision) Language Models (Sriram Balasubramanian, 2025) View paper
  - [35] Towards faithful chain-of-thought: Large language models are bridging reasoners (J Li, 2024) View paper
  - Explainability and Interpretability Perspectives (2 papers)
  - [19] Threading the Needle: Reweaving Chain-of-Thought Reasoning to Explain Human Label Variation (Beiduo Chen, 2025) View paper
  - [36] On Learning Verifiers for Chain-of-Thought Reasoning (Balcan, 2025) View paper
- Specialized Reasoning Tasks and Benchmarks
  - Multi-Turn and Interactive Reasoning (1 papers)
  - [26] TurnBench-MS: A Benchmark for Evaluating Multi-Turn, Multi-Step Reasoning in Large Language Models (Yiran Zhang, 2025) View paper
  - Multimodal and Visual Reasoning (3 papers)
  - [29] Sci-reason: A dataset with chain-of-thought rationales for complex multimodal reasoning in academic areas (MA Chenghao, 2025) View paper
  - [40] Seeing is Not Reasoning: MVPBench for Graph-based Evaluation of Multi-path Visual Physical CoT (Yi Junchao, 2025) View paper
  - [48] Journeybench: A challenging one-stop vision-language understanding benchmark of generated images (Hani Al-Omari, 2024) View paper
  - Domain-Specific Reasoning Applications (4 papers)
  - [25] Will trump win in 2024? predicting the us presidential election via multi-step reasoning with large language models (Chenxiao Yu, 2024) View paper
  - [38] A multi-step linguistic validation for cultural adaptation of the German-language Postpartum Bonding Questionnaire (Pia-Cecilia Steinbrueck, 2023) View paper
  - [46] Where Do We Go From Here? Multi-scale Allocentric Relational Inferencefrom Natural Spatial Descriptions (Kulkarni Sayali, 2024) View paper
  - [50] ASCENT (Automated Simulations to Characterize Electrical Nerve Thresholds): A pipeline for sample-specific computational modeling of electrical stimulation of â¦ (ED Musselman, 2021) View paper

Narrative

Core task: instance-level faithfulness detection of chain-of-thought reasoning. The field has organized itself around five major branches that reflect different facets of ensuring and understanding reasoning quality. Faithfulness Measurement and Detection Methods focus on diagnosing whether intermediate reasoning steps genuinely support final answers, often through causal interventions (Frit Causal Importance[4]) or symbolic verification (Symbolic Chain-of-Thought[1]). Verification and Correctness Assessment emphasizes automated checking mechanisms, including process reward models (Process Reward Models[5]) and deductive approaches (Deductive Verification[2]). Reasoning Improvement and Training Methods explore how to refine models through supervision signals and synthetic data generation. Analysis and Understanding of CoT Reasoning investigates the internal mechanisms and biases that shape reasoning behavior (Mechanistic Interpretation Multi-Step[14], Bias CoT Faithfulness[34]). Finally, Specialized Reasoning Tasks and Benchmarks provide domain-specific testbeds and evaluation protocols to stress-test reasoning capabilities across diverse problem settings.

A particularly active tension exists between comprehensive evaluation frameworks and targeted intervention studies. Works like Difficulty Faithful CoT[15] and Reflection Effectiveness Faithfulness[21] examine how task difficulty and self-correction influence faithfulness, revealing that harder problems often expose fragility in reasoning chains. FaithCoT-Bench[0] situates itself within the benchmarking strand, offering a systematic evaluation protocol that complements these neighboring studies by providing standardized metrics for instance-level detection. While Difficulty Faithful CoT[15] emphasizes the relationship between problem complexity and reasoning reliability, and Reflection Effectiveness Faithfulness[21] probes whether models can self-diagnose errors, FaithCoT-Bench[0] provides the infrastructure to measure these phenomena at scale. This positioning reflects a broader shift toward rigorous, reproducible assessment of faithfulness properties, bridging the gap between theoretical understanding of reasoning failures and practical detection tools that can guide model development and deployment decisions.

Related Works in Same Category

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

1. On the difficulty of faithful chain-of-thought reasoning in large language models

Authors: SH Tanneru, D Ley, C Agarwal | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â¦ faithfulness, as evidenced by a rise in faithfulness compared to the uniform counterpart, ie, â¦ In this study, we investigated the challenge of eliciting faithfulness chain-of-thought reasoning â¦

Relationship Analysis

Both papers belong to the Comprehensive Benchmarking and Evaluation Frameworks category, focusing on systematically evaluating instance-level CoT faithfulness. The original paper (FaithCoT-Bench) introduces a unified benchmark with expert-annotated dataset (FINE-COT) and evaluates 11 detection methods across multiple paradigms, while the candidate paper focuses on the difficulty of eliciting faithful CoT reasoning by evaluating three intervention approaches (ICL, fine-tuning, activation editing) to improve faithfulness. The key difference is that the original paper establishes a benchmark for detecting unfaithfulness with ground-truth annotations, whereas the candidate paper investigates methods to enhance faithfulness through model interventions, finding limited success across all approaches.

2. Towards better chain-of-thought: A reflection on effectiveness and faithfulness

Authors: Jiachun Li, Pengfei Cao, Yu-Bo Chen, Yubo Chen, Jiexin Xu, et al. (10 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Chain-of-thought (CoT) prompting demonstrates varying performance under different reasoning tasks. Previous work attempts to evaluate it but falls short in providing an in-depth analysis of patterns that influence the CoT. In this paper, we study the CoT performance from the perspective of effectiveness and faithfulness. For the former, we identify key factors that influence CoT effectiveness on performance improvement, including problem difficulty, information gain, and information flow. For th...

Relationship Analysis

Both papers belong to the Comprehensive Benchmarking and Evaluation Frameworks category, focusing on systematic evaluation of CoT faithfulness. They overlap in addressing instance-level faithfulness detection through multi-faceted evaluation approaches, including counterfactual methods, logit-based analysis, and LLM-as-judge paradigms. However, the original paper (FaithCoT-Bench) provides a unified benchmark with expert-annotated ground truth (FINE-COT dataset) and evaluates 11 existing detection methods, while the candidate paper focuses on analyzing effectiveness and faithfulness factors (problem difficulty, information gain, information flow) and proposes a novel mitigation algorithm (QUIRE) rather than establishing a comprehensive benchmark framework.

Contributions Analysis

Overall novelty summary. The paper introduces FaithCoT-Bench, a unified benchmark for instance-level CoT unfaithfulness detection, alongside FINE-CoT, an expert-annotated dataset of over 1,000 trajectories with fine-grained annotations. It resides in the 'Comprehensive Benchmarking and Evaluation Frameworks' leaf, which contains only three papers total within the broader 'Faithfulness Measurement and Detection Methods' branch. This represents a relatively sparse research direction within a 50-paper taxonomy, suggesting that systematic benchmarking infrastructure for instance-level faithfulness detection remains underdeveloped compared to other aspects of CoT reasoning quality.

The taxonomy reveals neighboring leaves focused on intervention-based detection (three papers using counterfactual methods), probabilistic guarantees (one paper), and mechanistic analysis (two papers). These sibling categories emphasize diagnostic techniques rather than evaluation infrastructure. The broader 'Verification and Correctness Assessment' branch (eleven papers across four leaves) addresses related but distinct concerns about answer correctness rather than reasoning faithfulness. The paper's benchmarking focus thus occupies a methodological niche: it provides evaluation protocols that complement but do not overlap with the causal intervention studies or mechanistic interpretability work in adjacent leaves.

Among 19 candidates examined through limited semantic search, none clearly refute the three main contributions. The benchmark contribution examined 10 candidates with no refutations; the annotated dataset examined 8 with none refuting; the systematic evaluation of eleven methods examined only 1 candidate. This search scope is modest relative to the field's breadth, and the absence of refutations may reflect both the limited search and the genuine scarcity of prior unified benchmarking efforts. The dataset contribution appears particularly distinctive given the emphasis on fine-grained, step-level annotations with expert labeling, though the small candidate pool examined limits confidence in this assessment.

Given the sparse population of the benchmarking leaf and the limited literature search scope, the work appears to address a recognized gap in evaluation infrastructure. However, the analysis covers only top-K semantic matches and does not exhaustively survey all faithfulness evaluation efforts. The contribution's novelty hinges partly on the integration of expert annotations with systematic method comparison, which neighboring papers in intervention-based or mechanistic categories do not emphasize. A more comprehensive search might reveal additional benchmarking efforts in adjacent communities or application domains.

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

Contribution 1: FaithCoT-Bench unified benchmark for instance-level CoT unfaithfulness detection

Description: The authors present FaithCoT-Bench, which integrates a rigorous task formulation that treats unfaithfulness detection as a discriminative decision problem, an expert-annotated dataset, and a systematic evaluation protocol into a single comprehensive framework for studying Chain-of-Thought faithfulness at the instance level.

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. Beyond Correctness: Rewarding Faithful Reasoning in Retrieval-Augmented Generation

URL: View paper

Brief Assessment

Rewarding Faithful Reasoning[52] focuses on faithfulness evaluation in retrieval-augmented generation with search agents, not on creating a unified benchmark for instance-level CoT unfaithfulness detection across general reasoning tasks.

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

URL: View paper

Brief Assessment

Monitoring Misaligned Reasoning[58] focuses on safety monitoring using CoT activations to predict alignment of final responses, not on creating a unified benchmark for instance-level unfaithfulness detection with expert annotations and systematic evaluation protocols.

3. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency

URL: View paper

Brief Assessment

MME-CoT[56] focuses on evaluating CoT reasoning quality, robustness, and efficiency in large multimodal models across six domains (math, science, OCR, logic, space-time, general scenes), not on detecting unfaithfulness at the instance level. The original paper addresses a fundamentally different problem: discriminative detection of whether individual CoT trajectories faithfully represent internal reasoning processes.

4. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning

URL: View paper

Brief Assessment

Making Reasoning Matter[53] focuses on measuring and improving faithfulness through manual evaluation of chain-of-thought quality, but does not present a unified benchmark framework with expert-annotated datasets and systematic evaluation protocols for instance-level unfaithfulness detection as described in the original paper's FaithCoT-Bench contribution.

5. Measuring faithfulness in chain-of-thought reasoning

URL: View paper

Brief Assessment

Measuring Faithfulness CoT[51] focuses on mechanism-level analysis through counterfactual interventions (adding mistakes, paraphrasing, early answering) to assess whether CoT reasoning is post-hoc or faithful in aggregate. It does not provide instance-level ground truth annotations, a discriminative task formulation, or expert-labeled datasets with fine-grained unfaithfulness causes and step-level evidence as the original paper does.

6. Auditing Meta-Cognitive Hallucinations in Reasoning Large Language Models

URL: View paper

Brief Assessment

Meta-Cognitive Hallucinations[54] focuses on auditing hallucination causality and chain disloyalty in reasoning chains, not on establishing a unified benchmark framework for instance-level unfaithfulness detection with expert annotations and systematic evaluation protocols.

7. Measuring chain of thought faithfulness by unlearning reasoning steps

URL: View paper

Brief Assessment

Unlearning Reasoning Steps[23] focuses on parametric faithfulness through machine unlearning interventions on model parameters, not on creating a unified benchmark with expert annotations and systematic evaluation protocols for instance-level detection as FaithCoT-Bench does.

8. Measuring the Faithfulness of Thinking Drafts in Large Reasoning Models

URL: View paper

Brief Assessment

Thinking Drafts Faithfulness[57] focuses on evaluating faithfulness in Large Reasoning Models (LRMs) with thinking drafts through counterfactual interventions, not on creating a unified benchmark with expert annotations for instance-level CoT unfaithfulness detection across diverse domains and models.

9. Measuring faithfulness of chains of thought by unlearning reasoning steps

URL: View paper

Brief Assessment

Unlearning Reasoning Steps[3] proposes a measurement framework (FUR) for parametric faithfulness through unlearning, not a unified benchmark with expert annotations and systematic evaluation protocols for instance-level detection.

10. How interpretable are reasoning explanations from prompting large language models?

URL: View paper

Brief Assessment

Interpretable Reasoning Explanations[55] focuses on multi-dimensional interpretability evaluation (faithfulness, robustness, utility) across prompting techniques, not on creating a unified benchmark for instance-level unfaithfulness detection with expert annotations and systematic evaluation protocols.

Contribution 2: FINE-COT expert-annotated dataset with fine-grained unfaithfulness annotations

Description: The authors construct FINE-COT, a dataset containing over 1,000 reasoning trajectories from four LLMs across four domains. Each trajectory is annotated by experts with faithfulness labels, fine-grained causes of unfaithfulness categorized into eight principles, and step-level evidence, providing ground truth for instance-level evaluation.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Exploring Chain-of-Thought Reasoning for Steerable Pluralistic Alignment

URL: View paper

Brief Assessment

Steerable Pluralistic Alignment[63] focuses on chain-of-thought reasoning for pluralistic value alignment and perspective-taking, not on constructing datasets with unfaithfulness annotations for reasoning trajectories.

2. FG-PRM: Fine-grained Hallucination Detection and Mitigation in Language Model Mathematical Reasoning

URL: View paper

Brief Assessment

FG-PRM[61] focuses on hallucination detection in mathematical reasoning with automated synthetic data generation, not expert-annotated unfaithfulness labels across diverse reasoning domains as in FINE-COT.

3. Are Machines Better at Slow Thinking? Unveiling Human-Machine Inference Gaps in Entailment Verification

URL: View paper

Brief Assessment

Slow Thinking Gaps[62] focuses on human-machine inference gaps in entailment verification tasks, not on annotating unfaithfulness in LLM reasoning trajectories. The candidate does not provide expert annotations of CoT unfaithfulness with fine-grained categorizations.

4. WikiDT: Visual-Based Table Recognition and Question Answering Dataset

URL: View paper

Brief Assessment

WikiDT[60] focuses on table-based visual question answering with hierarchical labels for table recognition and QA tasks, not on reasoning trajectory faithfulness or chain-of-thought evaluation. The datasets address fundamentally different problems in different domains.

5. A generalist medical language model for disease diagnosis assistance

URL: View paper

Brief Assessment

Generalist Medical Model[59] focuses on medical diagnosis with expert-annotated diagnostic data, not on reasoning trajectory faithfulness evaluation. The candidate's annotations address medical content quality, not chain-of-thought unfaithfulness patterns.

6. Evaluating Faithfulness in Agentic RAG Systems for e-Governance Applications Using LLM-Based Judging Frameworks

URL: View paper

Brief Assessment

Agentic RAG Faithfulness[64] focuses on faithfulness evaluation in RAG systems for e-governance applications, not on constructing expert-annotated datasets with fine-grained unfaithfulness annotations for reasoning trajectories across multiple domains and LLMs.

7. Towards Trustworthy AI: Frameworks for Evaluating Consistency in Language Models

URL: View paper

Brief Assessment

Consistency Evaluation Frameworks[65] appears to be a thesis proposal focused on evaluating consistency in language models. The provided context contains only the title page and certificate, with no technical content about datasets, annotations, or reasoning trajectories that could refute the ORIGINAL paper's contribution of constructing FINE-COT with expert annotations for unfaithfulness detection.

8. Towards Scalable Domain-Specific Document Annotation: A Semantic Archetype-Driven Framework

URL: View paper

Brief Assessment

Semantic Archetype Framework[66] focuses on domain-specific document annotation using semantic archetypes, not on annotating reasoning trajectories for faithfulness evaluation. The candidate addresses a completely different problem domain (document annotation) compared to the original's focus on chain-of-thought faithfulness detection.

Contribution 3: Systematic evaluation of eleven CoT faithfulness detection methods

Description: The authors perform a comprehensive benchmarking of eleven detection methods across three paradigms (counterfactual, logit-based, and LLM-as-judge), deriving empirical insights about their strengths, weaknesses, and performance variations across domains and model types.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Inducing Faithfulness in Structured Reasoning via Counterfactual Sensitivity

URL: View paper

Brief Assessment

Counterfactual Sensitivity Faithfulness[67] focuses on training-time regularization to induce faithfulness in reasoning, not on evaluating existing detection methods. The candidate introduces a training objective (CSR) rather than benchmarking detection approaches.

Appendix: Text Similarity Detection

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

References

- [0] FaithCoT-Bench: Benchmarking Instance-Level Faithfulness of Chain-of-Thought Reasoning View paper
- [1] Faithful logical reasoning via symbolic chain-of-thought View paper
- [2] Deductive Verification of Chain-of-Thought Reasoning View paper
- [3] Measuring faithfulness of chains of thought by unlearning reasoning steps View paper
- [4] Frit: Using causal importance to improve chain-of-thought faithfulness View paper
- [5] The Lessons of Developing Process Reward Models in Mathematical Reasoning View paper
- [6] Improve Mathematical Reasoning in Language Models by Automated Process Supervision View paper
- [7] Zero-Shot Verification-guided Chain of Thoughts View paper
- [8] ConCISE: Confidence-guided Compression in Step-by-step Efficient Reasoning View paper

- [9] Probabilistic soundness guarantees in llm reasoning chains View paper
- [10] Question decomposition improves the faithfulness of model-generated reasoning View paper
- [11] Hard2Verify: A Step-Level Verification Benchmark for Open-Ended Frontier Math View paper
- [12] Step-Wise Formal Verification for LLM-Based Mathematical Problem Solving View paper
- [13] RL on Incorrect Synthetic Data Scales the Efficiency of LLM Math Reasoning by Eight-Fold View paper
- [14] Towards a Mechanistic Interpretation of Multi-Step Reasoning Capabilities of Language Models View paper
- [15] On the difficulty of faithful chain-of-thought reasoning in large language models View paper
- [16] Stepwise Verification and Remediation of Student Reasoning Errors with Large Language Model Tutors View paper
- [17] LegalReasoner: Step-wised Verification-Correction for Legal Judgment Reasoning View paper
- [18] Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning View paper
- [19] Threading the Needle: Reweaving Chain-of-Thought Reasoning to Explain Human Label Variation View paper
- [20] Specializing smaller language models towards multi-step reasoning View paper
- [21] Towards better chain-of-thought: A reflection on effectiveness and faithfulness View paper
- [22] VerifiAgent: a Unified Verification Agent in Language Model Reasoning View paper
- [23] Measuring chain of thought faithfulness by unlearning reasoning steps View paper
- [24] Multi-step problem solving through a verifier: An empirical analysis on model-induced process supervision View paper
- [25] Will trump win in 2024? predicting the us presidential election via multi-step reasoning with large language models View paper
- [26] TurnBench-MS: A Benchmark for Evaluating Multi-Turn, Multi-Step Reasoning in Large Language Models View paper
- [27] R1-compress: Long chain-of-thought compression via chunk compression and search View paper
- [28] Chain-of-thought is not explainability View paper
- [29] Sci-reason: A dataset with chain-of-thought rationales for complex multimodal reasoning in academic areas View paper
- [30] Selfcheck: Using llms to zero-shot check their own step-by-step reasoning View paper
- [31] RelCheck: Improving Relation Extraction with Ontology-Guided and LLM-Based Validation View paper
- [32] R3-RAG: Learning Step-by-Step Reasoning and Retrieval for LLMs via Reinforcement Learning View paper
- [33] HoarePrompt: Structural Reasoning About Program Correctness in Natural Language View paper
- [34] A Closer Look at Bias and Chain-of-Thought Faithfulness of Large (Vision) Language Models View paper
- [35] Towards faithful chain-of-thought: Large language models are bridging reasoners View paper
- [36] On Learning Verifiers for Chain-of-Thought Reasoning View paper
- [37] Token Signature: Predicting Chain-of-Thought Gains with Token Decoding Feature in Large Language Models View paper
- [38] A multi-step linguistic validation for cultural adaptation of the German-language Postpartum Bonding Questionnaire View paper
- [39] RL Tango: Reinforcing Generator and Verifier Together for Language Reasoning View paper
- [40] Seeing is Not Reasoning: MVPBench for Graph-based Evaluation of Multi-path Visual Physical CoT View paper
- [41] Rex-Thinker: Grounded Object Referring via Chain-of-Thought Reasoning View paper
- [42] A mechanistic analysis of a transformer trained on a symbolic multi-step reasoning task View paper
- [43] SGEU: enhancing LLM reasoning via backward exemplar generation and verification: Z. Wang et al. View paper
- [44] On the Impact of Fine-Tuning on Chain-of-Thought Reasoning View paper
- [45] Faithful Chain-of-Thought Reasoning View paper
- [46] Where Do We Go From Here? Multi-scale Allocentric Relational Inferencefrom Natural Spatial Descriptions View paper
- [47] Improving LLM Reasoning through Scaling Inference Computation with Collaborative Verification View paper
- [48] Journeybench: A challenging one-stop vision-language understanding benchmark of generated images View paper
- [49] LaRS: Latent Reasoning Skills for Chain-of-Thought Reasoning View paper
- [50] ASCENT (Automated Simulations to Characterize Electrical Nerve Thresholds): A pipeline for sample-specific computational modeling of electrical stimulation of â¦ View paper
- [51] Measuring faithfulness in chain-of-thought reasoning View paper
- [52] Beyond Correctness: Rewarding Faithful Reasoning in Retrieval-Augmented Generation View paper
- [53] Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning View paper
- [54] Auditing Meta-Cognitive Hallucinations in Reasoning Large Language Models View paper
- [55] How interpretable are reasoning explanations from prompting large language models? View paper
- [56] Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency View paper
- [57] Measuring the Faithfulness of Thinking Drafts in Large Reasoning Models View paper
- [58] Can we predict alignment before models finish thinking? towards monitoring misaligned reasoning models View paper
- [59] A generalist medical language model for disease diagnosis assistance View paper
- [60] WikiDT: Visual-Based Table Recognition and Question Answering Dataset View paper
- [61] FG-PRM: Fine-grained Hallucination Detection and Mitigation in Language Model Mathematical Reasoning View paper
- [62] Are Machines Better at Slow Thinking? Unveiling Human-Machine Inference Gaps in Entailment Verification View paper
- [63] Exploring Chain-of-Thought Reasoning for Steerable Pluralistic Alignment View paper
- [64] Evaluating Faithfulness in Agentic RAG Systems for e-Governance Applications Using LLM-Based Judging Frameworks View paper
- [65] Towards Trustworthy AI: Frameworks for Evaluating Consistency in Language Models View paper
- [66] Towards Scalable Domain-Specific Document Annotation: A Semantic Archetype-Driven Framework View paper
- [67] Inducing Faithfulness in Structured Reasoning via Counterfactual Sensitivity View paper