Novelty Assessment Report

Paper: Verifying Chain-of-Thought Reasoning via its Computational Graph
PDF URL: https://openreview.net/pdf?id=CxiNICq0Rr
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

Current Chain-of-Thought (CoT) verification methods predict reasoning correctness based on outputs (black-box) or activations (gray-box), but offer limited insight into \textit{why} a computation fails. We introduce a white-box method: \textbf{Circuit-based Reasoning Verification (CRV)}. We hypothesize that attribution graphs of correct CoT steps, viewed as \textit{execution traces} of the model's latent reasoning circuits, possess distinct structural fingerprints from those of incorrect steps. By training a classifier on structural features of these graphs, we show that these traces contain a powerful signal of reasoning errors. Our white-box approach yields novel scientific insights unattainable by other methods. (1) We demonstrate that structural signatures of error are highly predictive, establishing the viability of verifying reasoning directly via its computational graph. (2) We find these signatures to be highly domain-specific, revealing that failures in different reasoning tasks manifest as distinct computational patterns. (3) We provide evidence that these signatures are not merely correlational; by using our analysis to guide targeted interventions on individual transcoder features, we successfully correct the model's faulty reasoning. Our work shows that, by scrutinizing a model's computational process, we can move from simple error detection to a deeper, causal understanding of LLM reasoning.

Core Task Landscape

This paper addresses: **Verifying Chain-of-Thought Reasoning Correctness via Computational Graph Analysis**

A total of **50 papers** were analyzed and organized into a taxonomy with **21 categories**.

Taxonomy Overview

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

- **Computational Graph and Circuit-Based Verification**
- **External Knowledge Graph Augmented Reasoning**
- **Graph-Structured Prompting and Reasoning Frameworks**
- **Verification and Validation via Output Analysis**
- **Reasoning Process Optimization and Training**
- **Compositional and Logical Reasoning Evaluation**
- **Interactive and Human-in-the-Loop Verification**
- **Specialized Domain Applications**
- **Theoretical Foundations and Synthesis**

Complete Taxonomy Tree

- Verifying Chain-of-Thought Reasoning Correctness via Computational Graph Analysis Survey Taxonomy
- Computational Graph and Circuit-Based Verification
  - Attribution Graph and Circuit Analysis ★ (3 papers)
  - [0] Verifying Chain-of-Thought Reasoning via its Computational Graph (Anon et al., 2026) View paper
  - [21] Mechanistic Unveiling of Transformer Circuits: Self-Influence as a Key to Model Reasoning (Zhang Lin, 2025) View paper
  - [44] Uncovering Graph Reasoning in Decoder-only Transformers with Circuit Tracing (Dai Xinnan, 2025) View paper
  - Structural Pattern Analysis in Reasoning Chains (2 papers)
  - [13] What characterizes effective reasoning? revisiting length, review, and structure of cot (Feng, 2025) View paper
  - [19] What Makes a Good Reasoning Chain? Uncovering Structural Patterns in Long Chain-of-Thought Reasoning (Jiang, 2025) View paper
- External Knowledge Graph Augmented Reasoning
  - KG-Guided Chain-of-Thought Generation (4 papers)
  - [10] Reasoning on graphs: Faithful and interpretable large language model reasoning (Luo, 2023) View paper
  - [11] Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph (Sun, 2023) View paper
  - [29] Graph elicitation for guiding multi-step reasoning in large language models (Park Jin-young, 2023) View paper
  - [43] Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs (Jin Bo-wen, 2024) View paper
  - Multi-Hop KG Reasoning and Retrieval (6 papers)
  - [23] DynaSearcher: Dynamic Knowledge Graph Augmented Search Agent via Multi-Reward Reinforcement Learning (Feng Wenfeng, 2025) View paper
  - [37] Graph-Augmented Reasoning: Evolving Step-by-Step Knowledge Graph Retrieval for LLM Reasoning (Wu Wenjie, 2025) View paper
  - [38] CogKR: Cognitive graph for multi-hop knowledge reasoning (Zhengxiao Du, 2021) View paper
  - [41] StepChain GraphRAG: Reasoning Over Knowledge Graphs for Multi-Hop Question Answering (Yuan Xin, 2025) View paper
  - [42] CHAINSFORMER: Numerical Reasoning on Knowledge Graphs from a Chain Perspective (Ze Zhao, 2025) View paper
  - [46] Dual view graph transformer networks for multi-hop knowledge graph reasoning. (Congcong Sun, 2025) View paper

    - KG Completion and Confidence Ranking (2 papers)
    - [4] GLR: Graph Chain-of-Thought with LoRA Fine-Tuning and Confidence Ranking for Knowledge Graph Completion (Yifei Chen, 2025) View paper
    - [27] Noisy positive-unlabeled learning with self-training for speculative knowledge graph reasoning (Wang, 2023) View paper
    - Domain-Specific KG Reasoning Applications (3 papers)
    - [32] Cognition-aware Knowledge Graph Reasoning for Explainable Recommendation (Qingyu Bing, 2023) View paper
    - [33] Structured reflective reasoning for precise medical knowledge graph retrieval augmented generation. (Beilun Wang, 2025) View paper
    - [47] A Graph-Based Computational Framework for Financial Statement Analysis Integrating BBRT Principles with Knowledge Graph Reasoning (Lu Gao, 2025) View paper
- Graph-Structured Prompting and Reasoning Frameworks
    - Graph-of-Thought and Structured Prompting (3 papers)
    - [3] Boosting logical reasoning in large language models through a new framework: The graph of thought (Lei Bin, 2023) View paper
    - [6] Structure guided prompt: Instructing large language model in multi-step reasoning by exploring graph structure of the text (Cheng Kewei, 2024) View paper
    - [26] Deliberate Reasoning in Language Models as Structure-Aware Planning with an Accurate World Model (Siheng Xiong, 2024) View paper
    - Agent-Based Graph Reasoning Systems (2 papers)
    - [5] Gptswarm: Language agents as optimizable graphs (M Zhuge, 2024) View paper
    - [8] Agentic Reasoning: A Streamlined Framework for Enhancing LLM Reasoning with Agentic Tools (Junde Wu, 2025) View paper
- Verification and Validation via Output Analysis
    - Graph-Based Output Verification (3 papers)
    - [7] Graphreason: Enhancing reasoning capabilities of large language models through a graph-based verification approach (Lang Cao, 2024) View paper
    - [16] Pelican: Correcting hallucination in vision-LLMs via claim decomposition and program of thought verification (Sahu, 2024) View paper
    - [24] Darg: Dynamic evaluation of large language models via adaptive reasoning graph (Jiaao Chen, 2024) View paper
    - Multi-Path and Structural Fact Verification (3 papers)
    - [14] GraphCheck: Multi-Path Fact-Checking with Entity-Relationship Graphs (Hye-Won Jeon, 2025) View paper
    - [15] Reasoning Paths as Signals: Augmenting Multi-hop Fact Verification through Structural Reasoning Progression (Zheng Liwen, 2025) View paper
    - [20] Reasoning over semantic-level graph for fact checking (Wanjun Zhong, 2020) View paper
    - Semantic and Discourse-Level Reasoning Verification (2 papers)
    - [31] Discourse-aware graph networks for textual logical reasoning (Yinya Huang, 2023) View paper
    - [36] ExplaGraphs: An explanation graph generation task for structured commonsense reasoning (Swarnadeep Saha, 2021) View paper
- Reasoning Process Optimization and Training
    - Process Reward Modeling and Reinforcement Learning (2 papers)
    - [2] Metastable Dynamics of Chain-of-Thought Reasoning: Provable Benefits of Search, RL and Distillation (Kim Juno, 2025) View paper
    - [12] Rewarding graph reasoning process makes llms more generalized reasoners (Miao Peng, 2025) View paper
    - Self-Training and Graph-Guided Learning (2 papers)
    - [9] STEP: Enhancing Video-LLMs' Compositional Reasoning by Spatio-Temporal Graph-guided Self-Training (Haiyi Qiu, 2025) View paper
    - [40] Reverse Chain-of-Thought and Causal Path Verification: A Modular Plugin for Aligning LLMs with Knowledge Graphs (Dezhuang Miao, 2025) View paper
- Compositional and Logical Reasoning Evaluation
    - Compositional Reasoning Limits and Benchmarks (2 papers)
    - [1] Faith and fate: Limits of transformers on compositionality (Dziri, 2023) View paper
    - [18] Clr-fact: Evaluating the complex logical reasoning capability of large language models over factual knowledge (Zheng, 2024) View paper
    - Code-Based and Algorithmic Reasoning (1 papers)
    - [22] GCoder: Improving Large Language Model for Generalized Graph Reasoning (Qifan Zhang, 2025) View paper
- Interactive and Human-in-the-Loop Verification (1 papers)
    - [50] Vis-CoT: A Human-in-the-Loop Framework for Interactive Visualization and Intervention in LLM Chain-of-Thought Reasoning (Kaviraj Pather, 2025) View paper
- Specialized Domain Applications
    - Software and Hardware Verification (5 papers)
    - [17] Hdreason: Algorithm-hardware codesign for hyperdimensional knowledge graph reasoning (Chen, 2024) View paper
    - [28] Less is more: Hop-wise graph attention for scalable and generalizable learning on circuits (Chenhui Deng, 2024) View paper
    - [30] A Multi-Agent Approach to Fault Localization via Graph-Based Retrieval and Reflexion (Md Nakhla Rafi, 2024) View paper
    - [45] Proving correctness of compilers using structured graphs (Patrick Bahr, 2014) View paper
    - [49] Software verification and graph similarity for automated evaluation of students' assignments (Milena VujoÅ¡eviÄ JaniÄiÄ, 2013) View paper
    - Security and Multi-Step Attack Reasoning (1 papers)
    - [39] Uncovering Multi-step Attacks with Threat Knowledge Graph Reasoning (Xiayu Xiang, 2024) View paper
    - Industrial Process and Task Planning (1 papers)
    - [35] A multi-step knowledge reasoning model based on information enhanced graph representation and deep learning model (Feng Li, 2024) View paper
    - Educational Assessment and Graph Comprehension (1 papers)
    - [34] Validating a measure of graph selection and graph reasoning for dynamic situations (Courtney Donovan, 2024) View paper
- Theoretical Foundations and Synthesis (2 papers)
    - [25] Toward a Mechanistic Understanding of Stepwise Inference in Transformers: A Synthetic Graph Navigation Model (M Khona, 2024) View paper

- [48] Graph-LLM fusion: enhancing fact representation and logical reasoning in artificial intelligence systems (Yang Juan, 2025) View paper

Narrative

Core task: Verifying chain-of-thought reasoning correctness via computational graph analysis. The field has evolved into a rich landscape organized around several complementary perspectives. At the highest level, one branch focuses on computational graph and circuit-based verification, examining how reasoning steps form analyzable structures and how internal model circuits can be inspected for correctness. Another major direction augments reasoning with external knowledge graphs, integrating structured world knowledge to ground and validate intermediate steps. Graph-structured prompting frameworks explore how to organize reasoning itself as a graph of interconnected thoughts, while verification via output analysis emphasizes post-hoc checking of generated reasoning chains. Additional branches address reasoning process optimization through training, compositional and logical evaluation benchmarks, interactive human-in-the-loop methods, specialized domain applications, and theoretical foundations. Works such as Graph of Thought[3] and Think on Graph[11] illustrate how graph representations can guide the reasoning process, while Faith and Fate[1] and GraphCheck[14] exemplify efforts to validate reasoning outputs.

Within this ecosystem, particularly active lines of work contrast mechanistic analysis of model internals with external validation strategies. Some studies like Mechanistic Unveiling[21] and Uncovering Graph Reasoning[44] probe the internal circuits and attribution graphs that underlie reasoning steps, seeking to understand what computational structures emerge during chain-of-thought generation. Others, such as GraphReason[7] and Reasoning on Graphs[10], leverage external knowledge graphs to anchor reasoning in verifiable facts. The original paper, Verifying CoT Graph[0], sits squarely within the computational graph and circuit-based verification branch, specifically focusing on attribution graph and circuit analysis. Its emphasis on analyzing the computational graph structure of reasoning chains aligns it closely with mechanistic approaches like Mechanistic Unveiling[21] and Uncovering Graph Reasoning[44], which similarly dissect internal reasoning pathways. This contrasts with works that primarily validate outputs against external references, positioning Verifying CoT Graph[0] as part of an emerging effort to make reasoning verification more intrinsic and interpretable through graph-theoretic analysis of the reasoning process itself.

Related Works in Same Category

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

1. Mechanistic Unveiling of Transformer Circuits: Self-Influence as a Key to Model Reasoning

Authors: Zhang Lin, Hu Lijie, Lin Zhang, Wang DI, Lijie Hu, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Transformer-based language models have achieved significant success; however, their internal mechanisms remain largely opaque due to the complexity of non-linear interactions and high-dimensional operations. While previous studies have demonstrated that these models implicitly embed reasoning trees, humans typically employ various distinct logical reasoning mechanisms to complete the same task. It is still unclear which multi-step reasoning mechanisms are used by language models to solve such ta...

Relationship Analysis

Both papers belong to the Attribution Graph and Circuit Analysis category, examining computational graphs and circuits within transformer models for verification purposes. They overlap in using attribution graphs to trace information flow through model components (attention heads, MLPs, and interpretable features) to understand reasoning processes. The key difference is that the original paper focuses on verifying CoT step correctness by training a classifier on structural graph features extracted from transcoder-based attribution graphs, while the candidate paper focuses on mechanistic interpretation of reasoning processes by calculating self-influence scores across circuit layers to map the model's thought process during inference.

---

2. Uncovering Graph Reasoning in Decoder-only Transformers with Circuit Tracing

Authors: Dai Xinnan, Xinnan Dai, Guo Kai, Chung-Hsiang Lo, Zeng, et al. (13 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Transformer-based LLMs demonstrate strong performance on graph reasoning tasks, yet their internal mechanisms remain underexplored. To uncover these reasoning process mechanisms in a fundamental and unified view, we set the basic decoder-only transformers and explain them using the circuit-tracer framework. Through this lens, we visualize reasoning traces and identify two core mechanisms in graph reasoning: token merging and structural memorization, which underlie both path reasoning and substru...

Relationship Analysis

Both papers belong to the Attribution Graph and Circuit Analysis category, examining computational graphs and circuit-based mechanisms in transformers for verification purposes. They overlap in using attribution graphs to trace information flow through model components and analyzing structural patterns in reasoning processes. However, the original paper focuses on verifying chain-of-thought reasoning correctness by extracting structural fingerprints from attribution graphs to classify step correctness, while the candidate paper investigates graph reasoning tasks (path finding, pattern extraction) to uncover token merging and structural memorization mechanisms through circuit visualization.

Contributions Analysis

Overall novelty summary. The paper introduces Circuit-based Reasoning Verification (CRV), which analyzes attribution graphs of chain-of-thought steps as execution traces of latent reasoning circuits. Within the taxonomy, it resides in the 'Attribution Graph and Circuit Analysis' leaf under 'Computational Graph and Circuit-Based Verification'. This leaf contains only three papers total, including the original work, indicating a relatively sparse and emerging research direction. The approach represents a white-box verification method that examines internal computational structures rather than relying solely on output analysis or external knowledge augmentation.

The taxonomy reveals that the broader field encompasses multiple verification paradigms. Neighboring branches include 'Structural Pattern Analysis in Reasoning Chains' (2 papers) within the same parent category, and more populated areas like 'External Knowledge Graph Augmented Reasoning' (15+ papers across multiple leaves) and 'Verification via Output Analysis' (8 papers). The scope note for the paper's leaf explicitly focuses on 'mechanistic circuits within transformer models', distinguishing it from methods that use external knowledge graphs or analyze only final outputs. This positioning suggests the work explores a less-traveled path compared to knowledge-graph-based or black-box verification approaches.

Among 30 candidates examined across three contributions, none were found to clearly refute any claimed novelty. For the core CRV method, 10 candidates were examined with 0 refutable overlaps; similarly, domain-specific structural signatures and causal interventions each had 10 candidates examined with no clear prior work. This limited search scope suggests that within the top-30 semantically similar papers, the specific combination of attribution graph analysis, structural fingerprinting of errors, and domain-specific patterns appears distinctive. However, the analysis acknowledges this represents a bounded literature search rather than exhaustive coverage.

Given the sparse population of the attribution graph analysis leaf and the absence of refuting work among examined candidates, the approach appears to occupy a relatively novel position within the limited search scope. The mechanistic focus on computational graph structures for verification contrasts with the field's heavier emphasis on external knowledge integration and output-based validation.

However, the analysis is constrained by examining only 30 candidates, leaving open the possibility of relevant work outside this semantic neighborhood.

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

Contribution 1: Circuit-based Reasoning Verification (CRV) method

Description: The authors propose CRV, a white-box verification method that analyzes the structural properties of attribution graphs constructed from interpretable transcoder features. By training a classifier on graph-based structural fingerprints, the method detects reasoning errors by examining the computational process rather than just outputs or raw activations.

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. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains

URL: View paper

Brief Assessment

Weakest Link[57] focuses on step-level annotation and verification of chain-of-thought reasoning chains using NLI-based methods and LM prompting, not on analyzing attribution graph structural properties or circuit-based computational analysis.

---

2. CiRLExplainer: Causality-Inspired Explainer for Graph Neural Networks via Reinforcement Learning

URL: View paper

Brief Assessment

CiRLExplainer[58] focuses on explaining GNN predictions through causal attribution and graph structure analysis, not on verifying chain-of-thought reasoning in language models or analyzing computational execution traces.

---

3. Beyond the Answer: Advancing Multi-Hop QA with Fine-Grained Graph Reasoning and Evaluation

URL: View paper

Brief Assessment

Fine Grained Graph[53] focuses on multi-hop QA evaluation using graph-structured plans for question decomposition and step-level answer verification, not on verifying chain-of-thought reasoning via attribution graph structural analysis of transcoder features.

---

4. KG-TRACES: Enhancing Large Language Models with Knowledge Graph-constrained Trajectory Reasoning and Attribution Supervision

URL: View paper

Brief Assessment

KG-TRACES[54] focuses on supervising LLMs to generate attribution-aware reasoning processes grounded in knowledge graph paths, not on analyzing attribution graph structural properties to detect reasoning errors. The candidate uses symbolic reasoning paths from KGs for supervision, while the original analyzes computational graph structures via transcoder features.

---

5. Is Sarcasm Detection A Step-by-Step Reasoning Process in Large Language Models?

URL: View paper

Brief Assessment

Sarcasm Detection Reasoning[52] focuses on sarcasm detection using prompting frameworks that elicit sequential and non-sequential reasoning cues, not on verifying chain-of-thought reasoning via computational graph analysis or attribution graph structural properties.

---

6. CoTAR: Chain-of-Thought Attribution Reasoning with Multi-level Granularity

URL: View paper

Brief Assessment

CoTAR[56] focuses on attribution-oriented question answering with chain-of-thought reasoning to improve citation accuracy, not on verifying reasoning correctness through attribution graph structural analysis. The methods and objectives are fundamentally different.

---

7. Reasoning Paths as Signals: Augmenting Multi-hop Fact Verification through Structural Reasoning Progression

URL: View paper

Brief Assessment

Reasoning Paths Signals[15] focuses on multi-hop fact verification using structured reasoning graphs for evidence retrieval and claim verification, not on analyzing attribution graphs of chain-of-thought reasoning steps to detect computational errors in LLMs.

---

8. RADAR: A Reasoning-Guided Attribution Framework for Explainable Visual Data Analysis

URL: View paper

Brief Assessment

RADAR[55] focuses on visual data analysis (charts/graphs) and attribution of reasoning to visual regions, not on verifying chain-of-thought reasoning via computational graph structural analysis. The candidate addresses a fundamentally different domain and task.

---

9. Graph elicitation for guiding multi-step reasoning in large language models

URL: View paper

Brief Assessment

Graph Elicitation[29] focuses on decomposing complex questions into sub-questions using knowledge triplets extracted from the question itself, not on verifying reasoning correctness through computational graph analysis. The candidate addresses question decomposition for multi-hop QA, while the original paper analyzes attribution graphs of model computations to detect reasoning errors.

---

10. Towards Faithful Multi-step Reasoning through Fine-Grained Causal-aware Attribution Reasoning Distillation

URL: View paper

Brief Assessment

The candidate paper (Faithful Multi Step[51]) is not available for comparison. Without access to its full text, I cannot assess whether it demonstrates prior work on attribution graph structural analysis for reasoning verification.

Contribution 2: Domain-specific structural signatures of reasoning errors

Description: The authors demonstrate through cross-domain experiments that error signatures in attribution graphs are task-specific. Failures in different reasoning domains (e.g., boolean logic versus arithmetic) produce distinct structural patterns, though a combined classifier can learn multiple failure geometries simultaneously.

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. Evaluating Tool Selection and Usage Efficiency of LLM-based Agents in Domain-Specific Tasks: A Comparative Analysis

URL: View paper

Brief Assessment

Tool Selection Efficiency[75] focuses on tool selection patterns and usage efficiency in LLM-based agents across different application domains (financial analysis, scientific computation, data processing), not on structural patterns of reasoning errors in computational graphs. The domains studied are application areas rather than reasoning task types (e.g., boolean logic vs. arithmetic).

---

2. Comprehension Without Competence: Architectural Limits of LLMs in Symbolic Computation and Reasoning

URL: View paper

Brief Assessment

Comprehension Without Competence[74] focuses on architectural limitations in symbolic computation and reasoning across arithmetic and relational domains, examining why LLMs fail at execution despite understanding principles. It does not address error signature analysis in attribution graphs or task-specific failure patterns in the computational graph structure that the original paper investigates.

---

3. Failuresensoriq: A multi-choice qa dataset for understanding sensor relationships and failure modes

URL: View paper

Brief Assessment

FailureSensorIQ[70] focuses on multi-choice QA for understanding sensor-failure relationships in industrial assets, not on analyzing computational graph structures or error patterns in reasoning tasks across different domains like boolean logic or arithmetic.

---

4. Engibench: A benchmark for evaluating large language models on engineering problem solving

URL: View paper

Brief Assessment

EngiBench[77] focuses on evaluating LLMs across hierarchical engineering problem-solving tasks (knowledge retrieval, contextual reasoning, open-ended modeling) rather than analyzing computational graph structures or attribution patterns in reasoning failures. The candidate does not address domain-specific error signatures in computational traces.

---

5. Stochastic subnetwork induction for contextual perturbation analysis in large language model architectures

URL: View paper

Brief Assessment

Stochastic Subnetwork[72] focuses on contextual perturbation analysis and subnetwork induction in LLM architectures. The provided context does not contain sufficient detail about domain-specific error patterns in reasoning tasks to challenge the original paper's novelty claim about task-specific failure geometries in attribution graphs.

---

6. Failure modes of llms for causal reasoning on narratives

URL: View paper

Brief Assessment

Failure Modes Causal[71] examines domain-specific failure patterns in causal reasoning on narratives (e.g., reliance on event ordering, parametric knowledge shortcuts), not computational graph structures in chain-of-thought reasoning. The domains and analytical approaches differ fundamentally.

---

7. FinEval-KR: A Financial Domain Evaluation Framework for Large Language Models' Knowledge and Reasoning

URL: View paper

Brief Assessment

FinEval KR[76] focuses on decoupling knowledge and reasoning abilities in financial domain evaluation using cognitive science frameworks, not on analyzing structural patterns in computational graphs across different reasoning domains like the original paper.

---

8. Art: Automatic multi-step reasoning and tool-use for large language models

URL: View paper

Brief Assessment

ART[69] focuses on automatic multi-step reasoning and tool-use for task decomposition, not on analyzing domain-specific computational patterns in reasoning failures. The candidate does not examine error signatures or structural patterns distinguishing correct from incorrect reasoning steps.

---

9. When thinking fails: The pitfalls of reasoning for instruction-following in llms

URL: View paper

Brief Assessment

Thinking Fails[78] focuses on instruction-following failures caused by reasoning (CoT prompting), not on domain-specific computational patterns in reasoning task failures. The candidate examines attention shifts and constraint violations in instruction-following, while the original analyzes structural signatures in attribution graphs across different reasoning domains (boolean logic vs. arithmetic).

---

10. Grammars of Formal Uncertainty: When to Trust LLMs in Automated Reasoning Tasks

URL: View paper

Brief Assessment

Grammars Formal Uncertainty[73] focuses on uncertainty quantification in LLM-generated formal artifacts (SMT-lib programs) using PCFGs, not on analyzing error signatures in attribution graphs across reasoning domains. The candidate's task-dependent uncertainty patterns relate to formalization quality, not computational graph structures in chain-of-thought reasoning.

Contribution 3: Causal interventions guided by mechanistic analysis

Description: The authors show that structural error signatures are causally implicated in reasoning failures by performing targeted interventions on specific transcoder features identified through their analysis. These interventions successfully correct computational errors, demonstrating that the method enables actionable model debugging beyond simple error detection.

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. Learning internal representations by error propagation

URL: View paper

Brief Assessment

Error Propagation[62] focuses on the backpropagation learning algorithm for neural networks, not on causal interventions to correct reasoning errors in language models. The candidate addresses a fundamentally different problem domain (training neural networks via gradient descent) rather than debugging and correcting computational errors in chain-of-thought reasoning.

---

2. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models

URL: View paper

Brief Assessment

CausalBench[61] focuses on evaluating causal reasoning capabilities through benchmark questions across text/math/code domains, not on performing targeted interventions on model features to correct computational errors during reasoning.

---

3. Towards Error Centric Intelligence I, Beyond Observational Learning

URL: View paper

Brief Assessment

Error Centric Intelligence[64] focuses on theoretical foundations for error-centric learning and structural principles (LAP, ICM, CAP) for general intelligence, not on empirical interventions on specific model features to correct reasoning errors in chain-of-thought tasks.

---

4. Causality-based neural network repair

URL: View paper

Brief Assessment

Causality Network Repair[63] focuses on repairing neural networks by identifying 'guilty' neurons through causality-based fault localization for properties like fairness and backdoor removal. The original paper performs causal interventions on transcoder features to correct reasoning errors in LLMs, which is a fundamentally different application domain and technical approach.

---

5. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks

URL: View paper

Brief Assessment

Transparent AI Survey[60] focuses on general interpretability methods for DNNs across various domains, not specifically on causal interventions for correcting reasoning errors in chain-of-thought processes as demonstrated in the original paper.

---

6. Coca: Improving and explaining graph neural network-based vulnerability detection systems

URL: View paper

Brief Assessment

Coca[68] focuses on explaining GNN-based vulnerability detection in code through causal inference to identify crucial statements, not on performing targeted interventions to correct computational errors in LLM reasoning chains.

---

7. Causality-Driven Neural Network Repair: Challenges and Opportunities

URL: View paper

Brief Assessment

Causality Driven Repair[65] focuses on general DNN repair through causal inference techniques for fairness and robustness, not on mechanistic interpretability of reasoning circuits or transcoder-based interventions on specific features identified through attribution graph analysis.

---

8. Causal intervention and parameter-free reasoning for few-shot SAR target recognition

URL: View paper

Brief Assessment

Causal Intervention SAR[67] applies causal intervention to SAR image features for target recognition, not to internal model features for correcting reasoning errors in language models.

---

9. Inference-time intervention: Eliciting truthful answers from a language model

URL: View paper

Brief Assessment

Inference Time Intervention[59] focuses on shifting activations along 'truthful' directions in attention heads to improve factual accuracy, not on correcting computational errors in chain-of-thought reasoning via transcoder features identified through structural error analysis.

---

10. Neural-Symbolic VideoQA: Learning Compositional Spatio-Temporal Reasoning for Real-world Video Question Answering

URL: View paper

Brief Assessment

Neural Symbolic VideoQA[66] focuses on compositional spatio-temporal reasoning for video question answering using symbolic representations and program executors. It does not perform causal interventions on model features to correct reasoning errors as described in the original paper's contribution.

Appendix: Text Similarity Detection

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

References

- [0] Verifying Chain-of-Thought Reasoning via its Computational Graph View paper
- [1] Faith and fate: Limits of transformers on compositionality View paper
- [2] Metastable Dynamics of Chain-of-Thought Reasoning: Provable Benefits of Search, RL and Distillation View paper
- [3] Boosting logical reasoning in large language models through a new framework: The graph of thought View paper
- [4] GLR: Graph Chain-of-Thought with LoRA Fine-Tuning and Confidence Ranking for Knowledge Graph Completion View paper
- [5] Gptswarm: Language agents as optimizable graphs View paper
- [6] Structure guided prompt: Instructing large language model in multi-step reasoning by exploring graph structure of the text View paper
- [7] Graphreason: Enhancing reasoning capabilities of large language models through a graph-based verification approach View paper
- [8] Agentic Reasoning: A Streamlined Framework for Enhancing LLM Reasoning with Agentic Tools View paper
- [9] STEP: Enhancing Video-LLMs' Compositional Reasoning by Spatio-Temporal Graph-guided Self-Training View paper
- [10] Reasoning on graphs: Faithful and interpretable large language model reasoning View paper
- [11] Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph View paper
- [12] Rewarding graph reasoning process makes llms more generalized reasoners View paper
- [13] What characterizes effective reasoning? revisiting length, review, and structure of cot View paper
- [14] GraphCheck: Multi-Path Fact-Checking with Entity-Relationship Graphs View paper
- [15] Reasoning Paths as Signals: Augmenting Multi-hop Fact Verification through Structural Reasoning Progression View paper
- [16] Pelican: Correcting hallucination in vision-LLMs via claim decomposition and program of thought verification View paper
- [17] Hdreason: Algorithm-hardware codesign for hyperdimensional knowledge graph reasoning View paper
- [18] Clr-fact: Evaluating the complex logical reasoning capability of large language models over factual knowledge View paper
- [19] What Makes a Good Reasoning Chain? Uncovering Structural Patterns in Long Chain-of-Thought Reasoning View paper
- [20] Reasoning over semantic-level graph for fact checking View paper
- [21] Mechanistic Unveiling of Transformer Circuits: Self-Influence as a Key to Model Reasoning View paper
- [22] GCoder: Improving Large Language Model for Generalized Graph Reasoning View paper
- [23] DynaSearcher: Dynamic Knowledge Graph Augmented Search Agent via Multi-Reward Reinforcement Learning View paper
- [24] Darg: Dynamic evaluation of large language models via adaptive reasoning graph View paper
- [25] Toward a Mechanistic Understanding of Stepwise Inference in Transformers: A Synthetic Graph Navigation Model View paper
- [26] Deliberate Reasoning in Language Models as Structure-Aware Planning with an Accurate World Model View paper
- [27] Noisy positive-unlabeled learning with self-training for speculative knowledge graph reasoning View paper
- [28] Less is more: Hop-wise graph attention for scalable and generalizable learning on circuits View paper
- [29] Graph elicitation for guiding multi-step reasoning in large language models View paper
- [30] A Multi-Agent Approach to Fault Localization via Graph-Based Retrieval and Reflexion View paper
- [31] Discourse-aware graph networks for textual logical reasoning View paper
- [32] Cognition-aware Knowledge Graph Reasoning for Explainable Recommendation View paper
- [33] Structured reflective reasoning for precise medical knowledge graph retrieval augmented generation. View paper
- [34] Validating a measure of graph selection and graph reasoning for dynamic situations View paper
- [35] A multi-step knowledge reasoning model based on information enhanced graph representation and deep learning model View paper
- [36] ExplaGraphs: An explanation graph generation task for structured commonsense reasoning View paper
- [37] Graph-Augmented Reasoning: Evolving Step-by-Step Knowledge Graph Retrieval for LLM Reasoning View paper
- [38] CogKR: Cognitive graph for multi-hop knowledge reasoning View paper
- [39] Uncovering Multi-step Attacks with Threat Knowledge Graph Reasoning View paper
- [40] Reverse Chain-of-Thought and Causal Path Verification: A Modular Plugin for Aligning LLMs with Knowledge Graphs View paper
- [41] StepChain GraphRAG: Reasoning Over Knowledge Graphs for Multi-Hop Question Answering View paper
- [42] CHAINSFORMER: Numerical Reasoning on Knowledge Graphs from a Chain Perspective View paper
- [43] Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs View paper
- [44] Uncovering Graph Reasoning in Decoder-only Transformers with Circuit Tracing View paper
- [45] Proving correctness of compilers using structured graphs View paper
- [46] Dual view graph transformer networks for multi-hop knowledge graph reasoning. View paper
- [47] A Graph-Based Computational Framework for Financial Statement Analysis Integrating BBRT Principles with Knowledge Graph Reasoning View paper
- [48] Graph-LLM fusion: enhancing fact representation and logical reasoning in artificial intelligence systems View paper
- [49] Software verification and graph similarity for automated evaluation of students' assignments View paper
- [50] Vis-CoT: A Human-in-the-Loop Framework for Interactive Visualization and Intervention in LLM Chain-of-Thought Reasoning View paper
- [51] Towards Faithful Multi-step Reasoning through Fine-Grained Causal-aware Attribution Reasoning Distillation View paper
- [52] Is Sarcasm Detection A Step-by-Step Reasoning Process in Large Language Models? View paper
- [53] Beyond the Answer: Advancing Multi-Hop QA with Fine-Grained Graph Reasoning and Evaluation View paper
- [54] KG-TRACES: Enhancing Large Language Models with Knowledge Graph-constrained Trajectory Reasoning and Attribution Supervision View paper
- [55] RADAR: A Reasoning-Guided Attribution Framework for Explainable Visual Data Analysis View paper
- [56] CoTAR: Chain-of-Thought Attribution Reasoning with Multi-level Granularity View paper
- [57] A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains View paper
- [58] CiRLExplainer: Causality-Inspired Explainer for Graph Neural Networks via Reinforcement Learning View paper
- [59] Inference-time intervention: Eliciting truthful answers from a language model View paper
- [60] Toward transparent ai: A survey on interpreting the inner structures of deep neural networks View paper
- [61] Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models View paper
- [62] Learning internal representations by error propagation View paper
- [63] Causality-based neural network repair View paper
- [64] Towards Error Centric Intelligence I, Beyond Observational Learning View paper
- [65] Causality-Driven Neural Network Repair: Challenges and Opportunities View paper

- [66] Neural-Symbolic VideoQA: Learning Compositional Spatio-Temporal Reasoning for Real-world Video Question Answering View paper
- [67] Causal intervention and parameter-free reasoning for few-shot SAR target recognition View paper
- [68] Coca: Improving and explaining graph neural network-based vulnerability detection systems View paper
- [69] Art: Automatic multi-step reasoning and tool-use for large language models View paper
- [70] Failuresensoriq: A multi-choice qa dataset for understanding sensor relationships and failure modes View paper
- [71] Failure modes of llms for causal reasoning on narratives View paper
- [72] Stochastic subnetwork induction for contextual perturbation analysis in large language model architectures View paper
- [73] Grammars of Formal Uncertainty: When to Trust LLMs in Automated Reasoning Tasks View paper
- [74] Comprehension Without Competence: Architectural Limits of LLMs in Symbolic Computation and Reasoning View paper
- [75] Evaluating Tool Selection and Usage Efficiency of LLM-based Agents in Domain-Specific Tasks: A Comparative Analysis View paper
- [76] FinEval-KR: A Financial Domain Evaluation Framework for Large Language Models' Knowledge and Reasoning View paper
- [77] Engibench: A benchmark for evaluating large language models on engineering problem solving View paper
- [78] When thinking fails: The pitfalls of reasoning for instruction-following in llms View paper