A Survey on Chain-of-Thought Reasoning Evaluation in Large Language Models: What, How, Who, and Where

Anonymous ACL submission

Abstract

As large language models evolve from conversational agents to reasoning engines, Chain-of-Thought (CoT) evaluation has become pivotal, yet remains fragmented and heavily reliant on final accuracy. This outcome bias often masks ***disguised accuracy***, failing to distinguish faithful reasoning from post-hoc rationalization. Despite the urgency, a comprehensive survey systematizing these process-oriented assessment techniques remains absent. To fill this gap, we present a unified framework organizing the literature along four dimensions: ***what to evaluate***, ***how to evaluate***, ***who evaluates***, and ***where to evaluate***. We formalize core quality metrics and systematically analyze evaluation methodologies across qualitative and quantitative paradigms within diverse application domains. By synthesizing these approaches to improve CoT reasoning reliability, this survey provides guidance for building robust evaluation pipelines and highlights key frontiers for future research.

1 Introduction

Recently, large language models (LLMs), such as GPT-5 (OpenAI, 2025) and Gemini-3 (Google DeepMind, 2025), have advanced toward sophisticated Chain-of-Thought (CoT) reasoning (Wei et al., 2022; Yang et al., 2025a; Guo et al., 2025a). This paradigm alleviates the traditional black-box nature of LLMs (Bommasani et al., 2021; Rae et al., 2021) by explicitly articulating intermediate reasoning steps akin to human cognition (Chu et al., 2024; Chen et al., 2025c). Such transparency not only enhances interpretability (Ye et al., 2023b; Chen et al., 2024a) but also improves performance across diverse tasks (Cobbe et al., 2021; Kojima et al., 2022; Zhang et al., 2023b; Nye et al., 2021; Wang et al., 2023c; Talmor et al., 2023).

However, as shown in Figure 1, this shift toward CoT reasoning introduces new challenges in assessing the reliability of the reasoning process

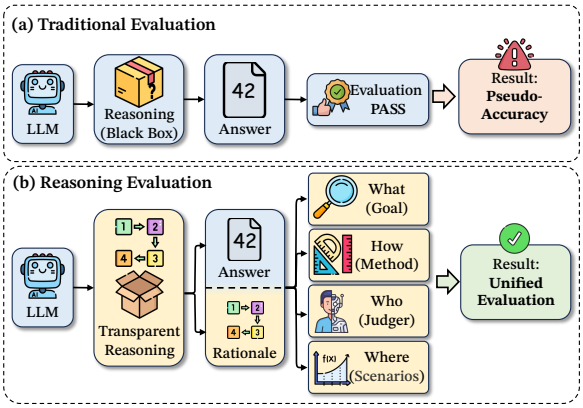


Figure 1: Unlike traditional outcome-based surveys, our survey systematically summarizes transparent CoT reasoning evaluation methodologies.

itself, beyond merely evaluating the correctness of final outputs (Madaan et al., 2023b; Lee and Hockenmaier, 2025; Zhao et al., 2025b). Conventional evaluation metrics, which primarily emphasize answer accuracy (Rajpurkar et al., 2016; Wang et al., 2018), fail to capture the nuanced quality of intermediate steps that underpin model behavior. Crucially, this ignores ***disguised accuracy*** [1], where models reach correct answers via flawed logic. Such outcome-bias leads to evaluations that fail to reflect the model's true capabilities.

Consequently, establishing reliable CoT evaluation is a central objective (Lee and Hockenmaier, 2025), essential for verifying reasoning soundness behind correct predictions and systematically characterizing LLM capabilities. Recent studies propose diverse approaches, from step-level assessments of process fidelity (Lightman et al., 2024; Uesato et al., 2022; Zhang et al., 2025d) to model-based judgment frameworks (Zheng et al., 2023; Kim et al., 2024b; Leang et al., 2025), to quantify such capabilities (Guan et al., 2025; Guo et al., 2025b). Despite these advances, current efforts

[1] An example is shown in Appendix A.

remain conceptually fragmented, and no comprehensive survey has yet unified evaluation principles across existing methodologies.

To fill this blank, we present the first comprehensive survey of CoT evaluation, organized under an integrative ***What-How-Who-Where*** structure. We begin by defining ***what to evaluate*** (§2), extending the notion of trustworthy reasoning beyond surface-level accuracy to encompass multiple cognitive and behavioral dimensions. We then classify ***how to evaluate*** (§3) across qualitative evaluation and quantitative evaluation, highlighting their methodological trade-offs and empirical foundations. Next, we examine ***who evaluates*** (§4), assessing the balance between human judgment and automated scalability in ensuring evaluation reliability. Finally, we map ***where to evaluate*** (§5) to application domains ranging from closed-world logical inference to open-ended dialogue and scientific reasoning. The concise taxonomy underpinning this framework is illustrated in Figure 2. Furthermore, we discuss emerging challenges and future directions (§6), aiming to guide the development of reliable CoT evaluation methodologies.

Our contributions are summarized as follows:

- **First Survey:** We present the first comprehensive survey on evaluating CoT reasoning.
- **Systematic Taxonomy:** We critically analyze current evaluation methodologies, identifying What-How-Who-Where dimensions in assessing the reliability of CoT reasoning.
- **New Frontiers:** We outline future research directions to advance the field of CoT evaluation, emphasizing the need for standardized benchmarks and more nuanced assessment criteria.

2 What to Evaluate

What qualities define a high-quality CoT?

Transcending outcome-based metrics that overlook reasoning nuances, Figure 3 illustrates our unified framework structured along five core dimensions.

2.1 Correctness

Correctness refers to the overall validity of CoT, capturing whether the final conclusion or result is factually and computationally sound (Wei et al., 2022; Kojima et al., 2022). It represents a foundational criterion for reasoning quality, as correctness establishes the necessary condition for any valid outcome (Cobbe et al., 2021; Uesato et al., 2022).

Formally, let a reasoning chain be defined as an ordered sequence of inferential steps $R = \{r_1, r_2, \ldots, r_n\}$, culminating in a final answer a. The reasoning chain satisfies overall correctness if and only if $\phi(a) = \texttt{True}$, where ϕ denotes a validation function that ensures factual, logical, and computational integrity.

2.2 Coherence

Coherence assesses logic structural soundness, ensuring deductive validity in reasoning irrespective of factual accuracy (Pan et al., 2023; Golovneva et al., 2023). Generally speaking, flawed logic can yield hallucinated conclusions (Zhang et al., 2024b), necessitating step-wise fidelity over mere overall correctness (Madaan et al., 2023b). Process reward models detect local non-sequiturs in verified chains (Lightman et al., 2024; Uesato et al., 2022). Symbolic solvers and logic checkers provide determinism and precision (Lyu et al., 2023a; Zhao et al., 2025b). Meanwhile, natural language inference models are utilized to measure step entailment and uncover cycles or disconnected reasoning (Zhao et al., 2025b; Valmeekam et al., 2022). Coherence thus anchors rational deduction, distinguishing it from narrative association.

For the reasoning steps $(r_i, r_{i+1}) \in R$, coherence is satisfied when entailment $r_i \models r_{i+1}$ holds for all i. Formally, logical coherence is defined as:

$$\mathcal{C}(R) = \frac{1}{n-1} \sum_{i=1}^{n-1} \texttt{Entail}(r_i, r_{i+1}), \quad (1)$$

where $\texttt{Entail}(\cdot)$ returns 1 if reasoning is valid.

2.3 Interpretability

Interpretability aligns model computations with human cognition by providing explanations that are not merely plausible but faithfully diagnostic (Guan et al., 2025; Qin et al., 2025). Attribution analysis quantifies token-level influence via $I(t_i) = \frac{\partial y}{\partial t_i}$ to reveal causal contributions (Lanham et al., 2023; Dziri et al., 2023). Complementarily, simulatability evaluates whether humans can approximate $f_\theta(x)$ from the reasoning trace R alone (Ye et al., 2023b; Golovneva et al., 2023). High interpretability thus enables epistemic alignment, rendering the model's latent reasoning cognitively reconstructable.

Formally, let $f_\theta(x) \rightarrow y$ denote the model mapping. Interpretability requires a human-comprehensible mapping $\Psi : R \mapsto E$, such that

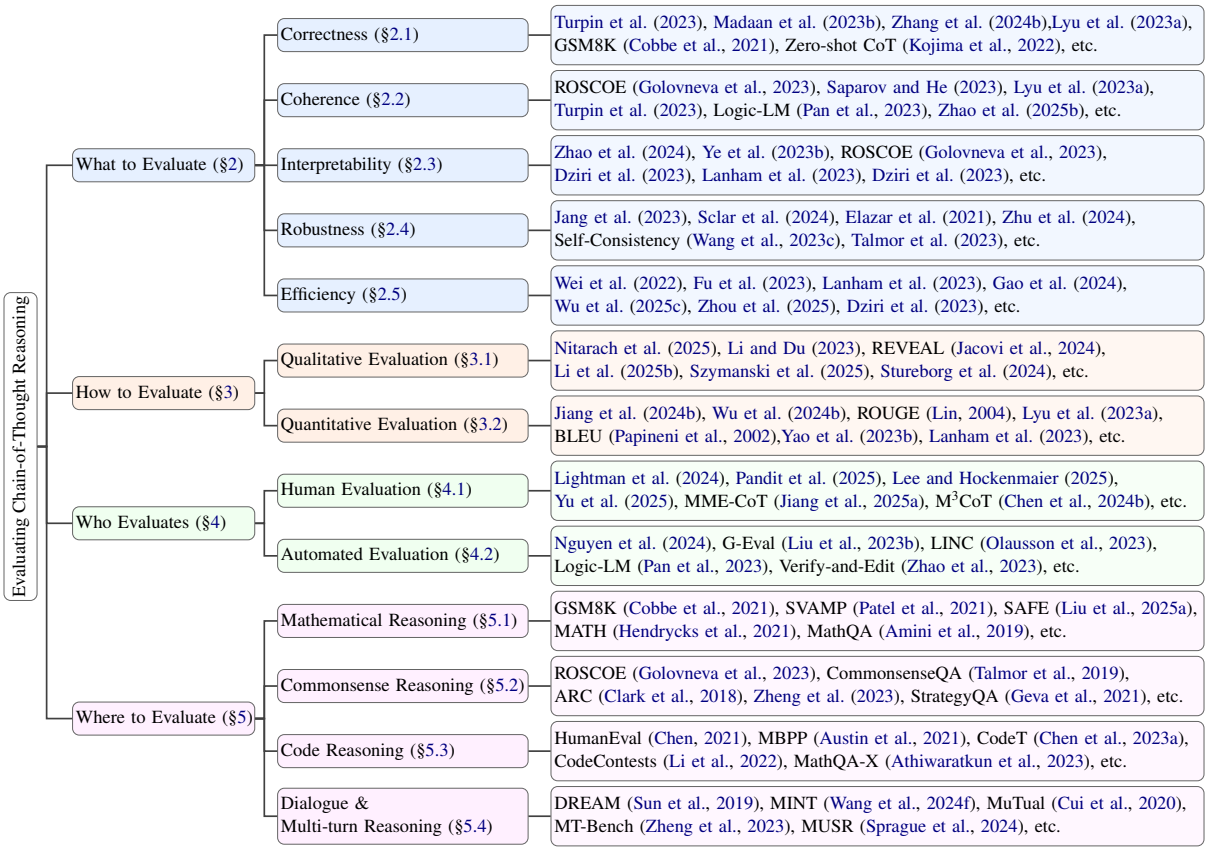


Figure 2: A brief taxonomy of representative works covered in our survey. The complete taxonomy is in Figure 6.

E faithfully captures the causal dependencies in f_θ. In other words, this formalizes the transparency and causal coherence of model reasoning.

2.4 Robustness

Robustness quantifies reasoning stability and reliability from random convergence by assessing the persistence of logical relations under paraphrasing or noise (Jang et al., 2023; Sclar et al., 2024). Typically, self-consistency metrics measure consensus across sampled reasoning paths (Wang et al., 2023c), while perturbation tests evaluate stability under adversarial or linguistic variations (Elazar et al., 2021; Zhu et al., 2024; Min et al., 2022).

For a model f_θ and perturbed input $x' \sim P(x)$, reasoning robustness is getting the most similar answer distributions:

$$\mathcal{R}(x) = 1 - \texttt{sim}(p(a|x), p(a|x')), \qquad (2)$$

where $\texttt{sim}$ is the similarity function of two probability distributions.

2.5 Efficiency

Efficiency reflects the optimal use of reasoning effort, maximizing progress per unit of computation. Practical assessment employs ratios of reasoning length to solution depth, penalizing unnecessary verbosity (Lanham et al., 2023; Wu et al., 2025c; Wang and Zhou, 2024). Efficient reasoning thus achieves maximal insight with minimal resources.

Given a reasoning sequence $R = \{r_1, \ldots, r_n\}$ with token cost $\tau(R)$ and performance or entropy $\mathcal{S}(R)$, the reasoning efficiency is quantified as

$$\mathcal{E}(R) = \frac{\mathcal{S}(R)}{\tau(R)}. \qquad (3)$$

3 How to Evaluate

How to evaluate the quality of CoT reasoning?

3.1 Qualitative Evaluation

3.1.1 Human Annotation

Human annotation provides gold-standard judgments for evaluating reasoning quality (Nitarach et al., 2025; Li and Du, 2023). To assess outputs systematically, define a fine-grained error taxonomy (e.g., factual errors, missing steps, and ambiguity) and verify each reasoning step to locate where CoT deviates from expert reasoning. Annotators also provide pairwise preferences (Zhang et al.,

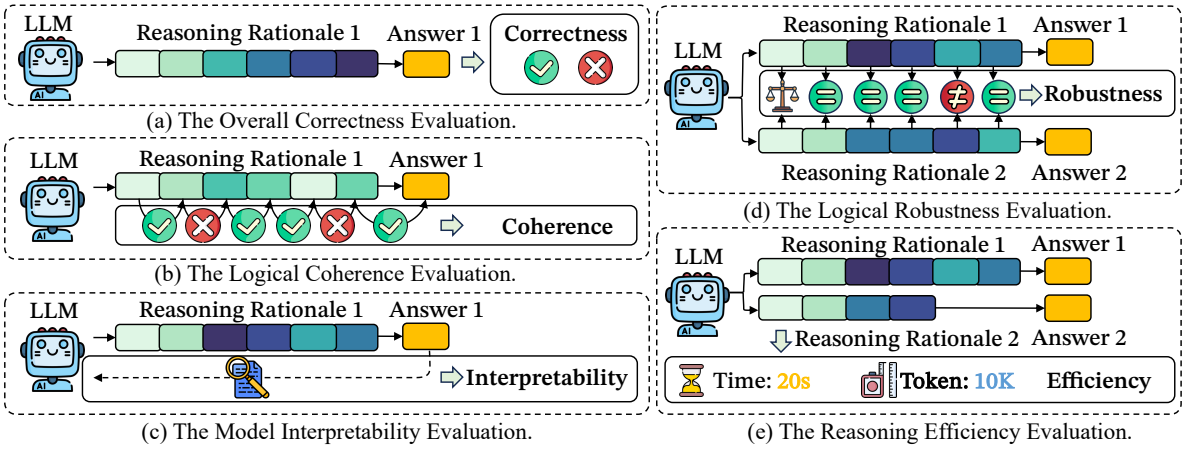


Figure 3: Five key dimensions of CoT evaluation: Correctness, Coherence, Interpretability, Robustness, and Efficiency. Each dimension captures a distinct facet of reasoning quality.

2024c; Zhao et al., 2025a; Yue et al., 2025), quantify agreement using Cohen's κ or Krippendorff's α and conduct periodic calibration to maintain consistency (Lee et al., 2024; Lyu et al., 2023b). Although costly to scale, this protocol yields reliable supervision for calibrating automated metrics and supporting robust model comparison (Chang et al., 2024; Chen et al., 2025e).

3.1.2 LLM-as-Judge

LLM-as-Judge uses advanced models to automate scalable reasoning assessment to provide a quality comment or select the superior reasoning trace between candidates, reducing calibration errors (Li et al., 2025b; Szymanski et al., 2025). For greater reliability, some frameworks augment qualitative critiques by CoT that anchor evaluations in explicit reasoning (Liu et al., 2023a; Jacovi et al., 2024). Step-level verification further breaks reasoning chains into atomic steps, checking each against formal constraints or external knowledge to detect hallucinations (Xu et al., 2024b; Paul et al., 2024). Despite the need for bias safeguards, LLM-as-Judge provides a scalable alternative to human experts for detailed reasoning evaluation.

3.2 Quantitative Evaluation

3.2.1 Correctness Evaluation

Quantitative metrics evaluate correctness by matching model outputs to gold-standard labels. For structured extraction, F1 and Exact Match test whether target tokens or error identifiers appear in the trace, enabling applications from pedagogical feedback (Jiang et al., 2024b) to debiasing interventions (Wu et al., 2024b). Recent studies use BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) to measure overlap between traces and reference explanations as a proxy for semantic similarity (Yao et al., 2023b; Maklad et al., 2025).

For subjective tasks, Likert ratings (1–5) score justification quality and are often combined with pairwise preferences against human reasoning (Bao et al., 2025; Jacovi et al., 2024; Li et al., 2025a). Discriminator-guided frameworks (Khalifa et al., 2023) and offline alignment systems (Wu et al., 2024a) train judge models to produce scores via ranking. Cog-CoT (Chen et al., 2025f) further uses hidden-state signals to predict trace reliability.

3.2.2 Coherence Evaluation

Intrinsic Evaluation assesses reasoning quality by separating the reasoning process from the final answer and focusing on structural and logical validity. It inspects intermediate steps for logical breaks, prioritizing derivation soundness over answer correctness (Golovneva et al., 2023; Lightman et al., 2024). Coherence is measured via step-wise analysis (Lyu et al., 2023a; Lanham et al., 2023), testing whether the reasoning chain causally supports the conclusion and separating deduction from heuristics (Saparov and He, 2023; Turpin et al., 2023). Building on this, structural analysis evaluates syntactic properties (e.g., step completeness and sequence length) to quantify inference validity and stability (Jiang et al., 2025b; Jin et al., 2024; Potamitis et al., 2025).

Recent methods capture reasoning integrity through logic-focused verification. VeriCoT (Feng et al., 2025) and DeduCE (Pandey et al., 2025) map language to symbolic logic for proof, while self-correction (Zhang et al., 2025a) and temporal checks (Mao et al., 2025) ensure consistency.

Causal metrics (Paul et al., 2024) and counterfactual tests (Wang et al., 2025b) distinguish inference from hallucinations. Granular tools include code unit tests (Saad-Falcon et al., 2024), logic probes (Kim et al., 2024a), and audits of multi-agent dynamics (Abdaljalil et al., 2025; She et al., 2025), shifting evaluation to structural audits.

Extrinsic Evaluation tests reasoning quality against external standards such as ground truth, structured knowledge, or execution environments (Lightman et al., 2024; Wang et al., 2024c; Liu et al., 2025a; Wang and Xu, 2025). This paradigm thus emphasizes objective verifiability rather than mere plausibility.

The most direct methods convert CoT into symbolic representations (Nguyen et al., 2025; Hu et al., 2025a,b) or executable code (Chen et al., 2022; Das et al., 2024), enabling deterministic checking (Olausson et al., 2023; Xu et al., 2024a; Feng et al., 2025; Simonds et al., 2025). LoT (Liu et al., 2025b) further integrate symbolic constraints dynamically into the reasoning process (Pan et al., 2023). Further, external compilers are applied to ensure language-agnostic verification (Nezhad et al., 2025). For knowledge-related reasoning, Graph-based methods (Luo et al., 2024; Amayuelas et al., 2025) trace reasoning paths through knowledge graphs, while fact decomposition (Min et al., 2023; Lage and Ostermann, 2025) and iterative feedback loops (Wang et al., 2024a) verify individual factual claims. For general generative tasks, semantic metrics (Sellam et al., 2020; Rei et al., 2020), like BERTScore (Zhang et al., 2020), quantify alignment with reference outputs, whereas structural metrics evaluate logical coherence (Prasad et al., 2023; Wang et al., 2025a; Hwang et al., 2025).

3.2.3 Robustness Evaluation

Robustness on Perturbed CoT steps. This robustness dimension tests whether models maintain reliable outputs under direct corruption of reasoning steps. Evaluations perturb intermediate CoT steps and measure effects on final predictions (Cui et al., 2025). Extending this, Yang et al. (2025b) apply structural variation tests that alter step syntax while preserving semantics, revealing how minor syntactic instabilities propagate to logical failures and exposing the reasoning chain as a vulnerability.

Robustness on CoT Reasoning Recovery. Complementarily, robustness evaluation measures the capacity to recover coherent reasoning by filtering or correcting noise in inputs or CoT traces. CD-CoT (Zhou et al., 2024) explicitly introduces inaccurate intermediate thoughts to assess self-correction capacity. Other protocols inject irrelevant context to test distraction resilience (Shi et al., 2023) or embed misleading opinions to measure sycophantic compliance, where the model aligns with incorrect user stances rather than factual correctness (Turpin et al., 2023).

Robustness across Multi-Samples. Beyond single trajectories, robustness can also arise from aggregating multiple stochastic reasoning paths. Self-consistency (Wang et al., 2023c) formalizes this idea by sampling diverse CoT traces and selecting the most consistent reasoning trace, with subsequent work enhancing this scheme through diversity- and robustness-oriented sampling (Li et al., 2023b) and adaptive stopping to reduce computation (Li et al., 2024d; Chen et al., 2023c; Lee et al., 2025a). Structural frameworks (Yao et al., 2023a; Besta et al., 2024) further generalize multi-sample reasoning into tree or graph topologies to more systematically explore robust solution spaces.

3.2.4 Interpretability Evaluation

Integration of Interpretability Methods. Recent studies integrate traditional explainability techniques to verify whether implicit token importance aligns with explicit CoT reasoning steps. Detailed methodologies, including gradient-based methods and counterfactual perturbations, appear in Appendix C. This approach distinguishes authentic reasoning from post-hoc rationalization.

Alignment with Human Preferences. Such evaluation measures how closely CoT structures match human patterns, beyond factual accuracy. The main method uses pairwise comparisons via crowdsourced platforms like ChatbotArena (Zheng et al., 2023), which yield Elo ratings from side-by-side rankings. To counter biases like verbosity preference (Gudibande et al., 2023), protocols apply process supervision: evaluators score individual intermediate steps for validity, separating logical soundness from style (Lightman et al., 2024; Uesato et al., 2022). For scalability, LLM-as-a-Judge (Wang et al., 2024g) mimic human rankings, but validation studies (Casper et al., 2023) stress calibration against expert judgments to detect subtle reasoning flaws (Chen et al., 2025e; Pather et al., 2025).

3.2.5 Efficiency Evaluation

To balance reasoning accuracy and computational overhead, evaluations quantify cognitive efficiency

via density metrics (Sui et al., 2025; Chen et al., 2025a). This entails calculating verbosity penalties and thought compression ratios (Saha et al., 2024; Ling et al., 2023b) to assess information density per token. Cost-effectiveness is evaluated by plotting accuracy-per-cost trade-offs, benchmarking frameworks like FrugalGPT (Chen et al., 2023b) and EcoAssistant (Zhang et al., 2023a) on dynamic query routing that invokes costly CoT paths only for complex cases, maximizing resource utility.

Takeaways

Prioritize Process: We should emphasize intrinsic validity and logical faithfulness to avoid "disguised accuracy", where correct answers conceal flawed reasoning.

Hybrid Evaluation: We should integrate human/LLM judges for semantic nuance with external solvers for rigorous symbolic verification.

Resilient Efficiency: We should maintain stability against adversarial noise while balancing reasoning depth and efficiency.

4 Who Evaluates

Who evaluates the quality of the CoT?

4.1 Human Evaluation

4.1.1 Expert Evaluation

Expert evaluation provides ground truth. For process supervision, Lightman et al. (2024) employ annotators to flag step-level errors. Some domains require rigorous verification: Pandit et al. (2025) recruit mathematicians, while legal experts assess complex reasoning (Lee and Hockenmaier, 2025; Yu et al., 2025). In multimodal settings (Jiang et al., 2025a; Chen et al., 2024b), humans check visual-textual alignment. However, Tutek et al. (2025) caution that such approval reflects plausibility rather than parametric faithfulness.

4.1.2 Crowdsourced Evaluation

Crowdsourced evaluation engages broad populations to align reasoning metrics with general intuition. Specifically, Golovneva et al. (2023) have workers validate error typologies, matching metrics like semantic consistency to layperson judgments. Likewise, Ramnath et al. (2024) gather large-scale feedback for MaRio's optimization, confirming preferences for plausible rationales. However, non-experts reduce precision. Kumar et al. (2025) show crowd workers miss subtle nuances versus domain experts, especially in complex tasks.

4.1.3 Human-In-The-Loop Evaluation

Human-In-The-Loop evaluation balances scalability and precision by directing human effort to key verification steps. Specifically, CRV (Zhao et al., 2025b) has humans validate model-flagged error fingerprints, avoiding full output review. To optimize effort, Diao et al. (2024) focus annotators on high-uncertainty samples in Active-Prompt, using human judgment to resolve ambiguity. Likewise, Li et al. (2023a) place humans as the final quality gate in HaluEval, with annotators conducting sanity checks only after programmatic filtering to minimize manual work.

4.2 Automated Evaluation

4.2.1 Symbolic and Rule-Based Verification

Symbolic and rule-based verification provides deterministic checks. For factuality, knowledge graphs anchor CoT to structured entities to evaluate hallucinations (Nguyen et al., 2024; Zhao et al., 2023; Peng et al., 2023). For logical rigor, solvers like LINC (Olausson et al., 2023), Logic-LM (Pan et al., 2023), and VeriCoT (Feng et al., 2025) convert steps to formal proofs, while theorem provers constrain generation to deductive rules (Poesia et al., 2023; Ye et al., 2023a; Mysore et al., 2023; Ling et al., 2023a).

4.2.2 Encoder-Based Verification

Encoder-based verification evaluates rationales via distance metrics against golden chains. While semantic encoders like BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020) measure similarity, they often overlook reasoning rigor. Addressing this, ROSCOE (Golovneva et al., 2023) assesses step-wise consistency, whereas RecEval (Prasad et al., 2023) employs natural language inference to verify strict logical entailment and mitigate shortcut learning.

4.2.3 Reward-Based Verification

Reward-based verification uses learned models to score reasoning quality (Lightman et al., 2024; Duan et al., 2025; Wang et al., 2025c). Process Reward Models (PRMs) act as step-by-step supervisors (Lightman et al., 2024; Wang et al., 2024c; Song et al., 2025), detecting intermediate errors overlooked by outcome metrics. Fine-grained detectors in REVEAL (Jacovi et al., 2024) and Li

et al. (2024b) categorize failures (e.g., calculation errors, symbol misuse), mimicking expert diagnostics. These PRMs evaluate logical transitions at each step (Uesato et al., 2022), reducing disguised accuracy (Bentham et al., 2024) and ensuring sound reasoning trajectories.

4.2.4 Generative Model Verification

Generative verification employs LLMs as judges to critique reasoning via semantic understanding. G-Eval (Liu et al., 2023b) uses GPT-4 for reference-free evaluation. Prompt-based systems like Self-Refine (Madaan et al., 2023b) and Self-Reflection (Shinn et al., 2023; Ji et al., 2023b; Weng et al., 2023) enable iterative self-correction of hallucinations. Recent benchmarks (Li et al., 2025a,b; Zhang et al., 2024c; Lee et al., 2025b; Chen et al., 2025c) assess these capabilities for expert-level rigor, while Zhang et al. (2025b) integrates verification into generation (Ali et al., 2025; Xu et al., 2025) to improve inference chains.

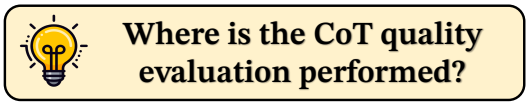
Takeaways

Automation Does Not Guarantee Objectivity: Automated methods are efficient but carry model biases and shortcuts.

Hybrid Intelligence as Next Paradigm: Human–AI evaluation is a distributed cognitive architecture. Future systems will exploit these complementarities.

Verification as Iterative Feedback: Evaluation evolves from post-hoc measurement to a core learning signal that refines alignment and reasoning fidelity.

5 Where to Evaluate



We review key evaluation aspects across common CoT scenarios here. Representative datasets and evaluation methodologies are in Appendix D.

5.1 Mathematical Reasoning

Mathematical reasoning failures range from calculation errors within coherent logic (e.g., Minerva (Lewkowycz et al., 2022)) to reasoning failures, such as invalid deductions or knowledge gaps, that undermine validity (Jiang et al., 2024b; Golovneva et al., 2023). Crucially, outcome accuracy often diverges from process validity, masking poor reasoning behind high scores (Guo et al., 2025b). While PRMs address this via consistency checks, their reliability remains limited by noisy reward estimation (Zhang et al., 2025d).

5.2 Commonsense Reasoning

Commonsense reasoning evaluation typically uses external verification against knowledge bases like ConceptNet (Liu and Singh, 2004) and ATOMIC (Sap et al., 2019). Methods always align intermediate steps with explicit knowledge paths, ensuring Correctness via factual grounding rather than plausibility alone (Feng et al., 2024; Wang et al., 2024a). However, quantifying logical coherence is challenging and still heavily relies on human judgment, as commonsense is probabilistic and allows multiple valid traces (Miao et al., 2024).

5.3 Code Reasoning

Code reasoning functions as a deterministic sandbox, verifying validity via execution (Chen, 2021). Evaluation should extend beyond correctness to capture efficiency and robustness. Benchmarks like APPS (Hendrycks et al., 2021) incorporate resource constraints, while CodeT (Chen et al., 2023a) employs dual-execution to reveal latent failures overlooked by standard checks.

5.4 Dialogue and Multi-turn Reasoning

In dialogue and multi-turn settings, logical coherence hinges on interpretability and reliable coreference resolution amid implicit assumptions. MuTual (Cui et al., 2020) stresses history-based inference, and QReCC (Anantha et al., 2021) assesses state tracking through context-dependent query rewriting. Coherence must persist across interactions: DNLI (Welleck et al., 2019) classifies response relations as entailment, neutral, or contradiction, while CICERO (Ghosal et al., 2022) penalizes hallucinations that disrupt causal continuity and conversational flow. [2]

Takeaways

Different Evaluation Scopes: Evaluation scenarios vary in CoT reasoning demands.

External Grounding: Validate reasoning with executable suites or knowledge graphs, beyond surface plausibility.

Contextual Consistency: In long-context scenarios, ensure reliability by tracking implicit assumptions and maintaining causal coherence across turns.

[2] Scientific reasoning is discussed in Appendix D.5.

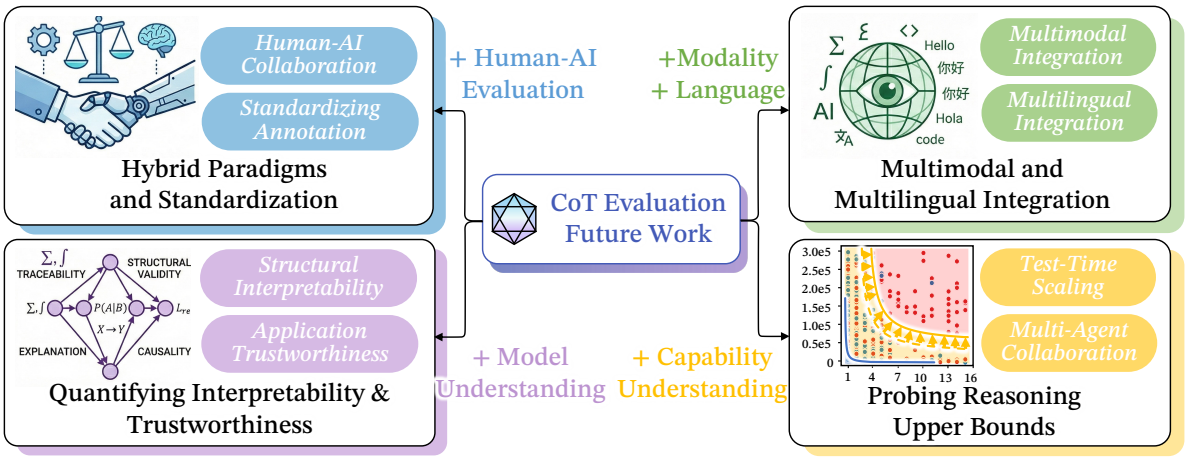


Figure 4: Future Directions for CoT Evaluation. We identify four key directions to advance the field.

6 Future Directions

As LLMs evolve from chat agents into specialized reasoning engines, CoT evaluation should transform in parallel to ensure rigorous assessment. We identify four critical frontiers (illustrated in Figure 4) essential to bridge current methodological gaps. For a comprehensive discussion of these strategic research trajectories, please refer to Appendix E.

6.1 Hybrid Paradigms and Standardization

Future protocols should prioritize **human-AI collaboration**, using models for initial filtering and humans for targeted verification (Diao et al., 2024; Li et al., 2023a). **Standardizing annotation** guidelines and testing adversarial robustness are crucial for scalable assessment (Cui et al., 2025). Thus, research must develop techniques to generate adversarial samples that stress-test metrics against subtle prompt variations.

6.2 Multimodal and Multilingual Frontiers

Evaluation should extend beyond text-only settings to multimodal models (Jiang et al., 2025a; Wu et al., 2025a) and **multilingual** CoT reasoning (Qin et al., 2023), with explicit tests of cross-modal grounding and cross-lingual logical consistency. To reduce Anglocentric bias and to better characterize **multimodal** capabilities, benchmarks should adopt more fine-grained protocols that cover diverse languages and require deeper visual–text reasoning (Bang et al., 2023; Chen et al., 2024b).

6.3 Quantifying Interpretability & Trustworthiness

Research should quantify **structural interpretability** by linking CoT traces to causal or knowledge graphs (Amayuelas et al., 2025; Nguyen et al., 2024) and using XAI to attribute input-context influence per reasoning step (Turpin et al., 2023). In law and healthcare, evaluations should prioritize **application trustworthiness** over accuracy (Yu et al., 2025). Enterprise frameworks should therefore enable interpretability, bias detection, and auditability (Shaikh et al., 2023).

6.4 Probing Reasoning Upper Bounds

Current benchmarks largely measure static inference, overlooking the potential of **test-time scaling** where models simulate deliberate thinking. In this paradigm, extra compute translates into verified gains through extended reasoning steps or deeper search (Chen et al., 2024a, 2025b; Snell et al., 2024; Muennighoff et al., 2025; Zhang et al., 2025c; Chen et al., 2025d). Similarly, **multi-agent collaboration** leverages ensemble diversity to estimate swarm-intelligence upper bounds (Du et al., 2023; Guo et al., 2024; Qian et al., 2024). Future evaluation should quantify the efficiency-performance trade-off, measuring how reasoning capabilities evolve as computing and collaboration scale.

7 Conclusion

This survey demonstrates that evaluating reasoning differs fundamentally from scoring outcomes. Transcending accuracy as a weak proxy, we structure assessment along five core dimensions to frame evaluation as a multifaceted design space. Reliable verification demands a synergy of symbolic rigor, scalable automation, and human judgment rather than a single paradigm. Future progress relies on multimodal, multilingual hybrids for trustworthy reasoning bounds. We hope this work offers insightful perspectives and foundational guidance for advancing the CoT reasoning evaluation field.

Limitations

Despite our efforts to provide a systematic and comprehensive survey, this work faces several limitations inherent to the rapid evolution of the field.

Scope Delimitation. To maintain depth, we strictly narrowed our scope to the *evaluation* of Chain-of-Thought reasoning. While training methodologies (e.g., Process Reward Models, RLHF) inherently involve evaluation steps, we excluded papers that focus solely on optimizing training objectives without introducing distinct assessment metrics or frameworks. This boundary, while necessary for focus, excludes broader alignment techniques that indirectly impact reasoning fidelity.

Taxonomical Fluidity. The proposed "What-How-Who-Where" framework serves as a structured lens to organize fragmented literature. In practice, however, these dimensions are not mutually exclusive. Many hybrid systems span multiple categories (e.g., a generative judge employing rule-based constraints), and our categorization reflects their primary methodological contribution rather than a rigid classification.

References

Samir Abdaljalil, Hasan Kurban, Khalid Qaraqe, and Erchin Serpedin. 2025. Theorem-of-thought: A multi-agent framework for abductive, deductive, and inductive reasoning in language models. *ArXiv preprint*, abs/2506.07106.

Nassir Jabir Al-Khafaji and Basit Khalaf Majeed. 2024. Evaluating large language models using arabic prompts to generate python codes. In *2024 4th International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, pages 1–5. IEEE.

Yusuf Ali, Gryphon Patlin, Karthik Kothuri, Muhammad Zubair Irshad, Wuwei Liang, and Zsolt Kira. 2025. Eve: A generator-verifier system for generative policies. *arXiv preprint arXiv:2512.21430*.

Aylton Almeida, Laerte Xavier, and Marco Tulio Valente. 2024. Automatic library migration using large language models: First results. In *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 427–433.

Alfonso Amayuelas, Joy Sain, Simerjot Kaur, and Charese Smiley. 2025. Grounding llm reasoning with knowledge graphs. *ArXiv preprint*, abs/2502.13247.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Ben Athiwaratkun, Sanjay Krishna Gouda, Zijian Wang, Xiaopeng Li, Yuchen Tian, Ming Tan, Wasi Uddin Ahmad, Shiqi Wang, Qing Sun, Mingyue Shang, Sujan Kumar Gonugondla, Hantian Ding, Varun Kumar, Nathan Fulton, Arash Farahani, Siddhartha Jain, Robert Giaquinto, Haifeng Qian, Murali Krishna Ramanathan, and Ramesh Nallapati. 2023. Multilingual evaluation of code generation models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and 1 others. 2021. Program synthesis with large language models. *ArXiv preprint*, abs/2108.07732.

Sriram Balasubramanian, Samyadeep Basu, and Soheil Feizi. 2025. A closer look at bias and chain-of-thought faithfulness of large (vision) language models. *ArXiv preprint*, abs/2505.23945.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Guangsheng Bao, Hongbo Zhang, Cunxiang Wang, Linyi Yang, and Yue Zhang. 2025. How likely do

llms with cot mimic human reasoning? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7831–7850.

Oliver Bentham, Nathan Stringham, and Ana Marasović. 2024. Chain-of-thought unfaithfulness as disguised accuracy. *ArXiv preprint*, abs/2402.14897.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17682–17690. AAAI Press.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, Ben He, Shanshan Jiang, and Bin Dong. 2024. ChatGPT is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3098–3110, Torino, Italia. ELRA and ICCL.

Rishi Bommasani and 1 others. 2021. On the opportunities and risks of foundation models. *ArXiv preprint*, abs/2108.07258.

Ana Brassard, Benjamin Heinzerling, Keito Kudo, Keisuke Sakaguchi, and Kentaro Inui. 2024. Acorn: Aspect-wise commonsense reasoning explanation evaluation. *ArXiv preprint*, abs/2405.04818.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, and 1 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *ArXiv preprint*, abs/2307.15217.

Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. clembench: Using game play to evaluate chat-optimized language models as conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11174–11219, Singapore. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. 2023a. Codet: Code generation with generated tests. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023b. Frugalgpt: How to use large language models while reducing cost and improving performance. *ArXiv preprint*, abs/2305.05176.

Mark Chen. 2021. Evaluating large language models trained on code. *ArXiv preprint*, abs/2107.03374.

Qiguang Chen, Dengyun Peng, Jinhao Liu, HuiKang Su, Jiannan Guan, Libo Qin, and Wanxiang Che. 2025a. Aware first, think less: Dynamic boundary self-awareness drives extreme reasoning efficiency in large language models. *arXiv preprint arXiv:2508.11582*.

Qiguang Chen, Libo Qin, Jinhao Liu, Yue Liao, Jiaqi Wang, Jingxuan Zhou, and Wanxiang Che. 2025b. Rbf++: Quantifying and optimizing reasoning boundaries across measurable and unmeasurable capabilities for chain-of-thought reasoning. *arXiv preprint arXiv:2505.13307*.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025c. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *ArXiv preprint*, abs/2503.09567.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiaqi Wang, Mengkang Hu, Zhi Chen, Wanxiang Che, and Ting Liu. 2025d. Ecm: A unified electronic circuit model for explaining the emergence of in-context learning and chain-of-thought in large language model. *arXiv preprint arXiv:2502.03325*.

Qiguang Chen, Libo Qin, Jiaqi Wang, Jingxuan Zhou, and Wanxiang Che. 2024a. Unlocking the capabilities of thought: A reasoning boundary framework to quantify and optimize chain-of-thought. *Advances in Neural Information Processing Systems*, 37:54872–54904.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024b. M^3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proc. of ACL*.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *ArXiv preprint*, abs/2211.12588.

Xiaoshu Chen, Sihang Zhou, Ke Liang, Duanyang Yuan, Haoyuan Chen, Xiaoyu Sun, Linyuan Meng, and Xinwang Liu. 2025e. Putting on the thinking hats: A survey on chain of thought fine-tuning from the perspective of human reasoning mechanism. *arXiv preprint arXiv:2510.13170*.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023c. Universal self-consistency for large language model generation. *ArXiv preprint*, abs/2311.17311.

Zhipeng Chen, Kun Zhou, Beichen Zhang, Zheng Gong, Xin Zhao, and Ji-Rong Wen. 2023d. ChatCoT: Tool-augmented chain-of-thought reasoning on chat-based large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14777–14790, Singapore. Association for Computational Linguistics.

Zijun Chen, Wenbo Hu, and Richang Hong. 2025f. Deep hidden cognition facilitates reliable chain-of-thought reasoning. *ArXiv preprint*, abs/2507.10007.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv preprint*, abs/1803.05457.

Karl Cobbe and 1 others. 2021. Training verifiers to solve math word problems. volume abs/2110.14168.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

Yu Cui, Bryan Hooi, Yujun Cai, and Yiwei Wang. 2025. Process or result? manipulated ending tokens can mislead reasoning llms to ignore the correct reasoning steps. *ArXiv preprint*, abs/2503.19326.

Debrup Das, Debopriyo Banerjee, Somak Aditya, and Ashish Kulkarni. 2024. MATHSENSEI: A tool-augmented large language model for mathematical reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 942–966, Mexico City, Mexico. Association for Computational Linguistics.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xun Deng, Sicheng Zhong, Honghua Dong, Jingyu Hu, Sidi Mohamed Beillahi, Xujie Si, and Fan Long. 2024. Assessing code generation with intermediate languages. *ArXiv preprint*, abs/2407.05411.

Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. 2024. Active prompting with chain-of-thought for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1350.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Keyu Duan, Zichen Liu, Xin Mao, Tianyu Pang, Changyu Chen, Qiguang Chen, Michael Qizhe Shieh, and Longxu Dou. 2025. Efficient process reward model training via active learning. *arXiv preprint arXiv:2504.10559*.

Zackary Okun Dunivin. 2024. Scalable qualitative coding with llms: Chain-of-thought reasoning matches human performance in some hermeneutic tasks. *ArXiv preprint*, abs/2401.15170.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaïd Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Run-Ze Fan, Zengzhi Wang, and Pengfei Liu. 2025. Megascience: Pushing the frontiers of post-training datasets for science reasoning. *ArXiv preprint*, abs/2507.16812.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*,

pages 1171–1182, Toronto, Canada. Association for Computational Linguistics.

Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. 2024. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *ArXiv preprint*, abs/2406.09098.

Yu Feng, Nathaniel Weir, Kaj Bostrom, Sam Bayless, Darion Cassel, Sapana Chaudhary, Benjamin Kiesl-Reiter, and Huzefa Rangwala. 2025. Vericot: Neuro-symbolic chain-of-thought validation via logical consistency checks. *ArXiv preprint*, abs/2511.04662.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Peng Gao and 1 others. 2024. Ruler: What's the real reasoning capability of your language model? *ArXiv preprint*, abs/2402.18679.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2022. CICERO: A dataset for contextualized commonsense inference in dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5010–5028, Dublin, Ireland. Association for Computational Linguistics.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Google DeepMind. 2025. Gemini 3 models and ecosystem. Online.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2023. Think before you speak: Training language models with pause tokens. *arXiv preprint arXiv:2310.02226*.

Alex Gu, Baptiste Rozière, Hugh James Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida Wang. 2024. Cruxeval: A benchmark for code reasoning, understanding and execution. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Melody Y. Guan, Miles Wang, Micah Carroll, Zehao Dou, Annie Y. Wei, Marcus Williams, Benjamin Arnav, Joost Huizinga, Ian Kivlichan, Mia Glaese, Jakub Pachocki, and Bowen Baker. 2025. Monitoring monitorability.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *ArXiv preprint*, abs/2305.15717.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv preprint*, abs/2501.12948.

Jiaxing Guo, Wenjie Yang, Shengzhong Zhang, Tongshan Xu, Lun Du, Da Zheng, and Zengfeng Huang. 2025b. Right is not enough: The pitfalls of outcome supervision in training llms for math reasoning. *ArXiv preprint*, abs/2506.06877.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, and 1 others. 2023. Evaluating large language models: A comprehensive survey. *ArXiv preprint*, abs/2310.19736.

Zhiyuan He and Dingmin Wang. 2025. When do symbolic solvers enhance reasoning in large language models? *ArXiv preprint*, abs/2512.03272.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *ArXiv preprint*, abs/2103.03874.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Md Sifat Hossain, Anika Tabassum, Md Fahim Arefin, and Tarannum Shaila Zaman. 2025. Llm-pros: Analyzing large language models' performance in competitive problem solving. In *2025 IEEE/ACM International Workshop on Large Language Models for Code (LLM4Code)*, pages 80–87. IEEE.

Mengkang Hu, Tianxing Chen, Yude Zou, Yuheng Lei, Qiguang Chen, Ming Li, Yao Mu, Hongyuan Zhang, Wenqi Shao, and Ping Luo. 2025a. Text2world:

Benchmarking large language models for symbolic world model generation. *arXiv preprint arXiv:2502.13092*.

Mengkang Hu, Bowei Xia, Yuran Wu, Ailing Yu, Yude Zou, Qiguang Chen, Shijian Wang, Jiarui Jin, Kexin Li, Wenxiang Jiao, and 1 others. 2025b. Agent2world: Learning to generate symbolic world models via adaptive multi-agent feedback. *arXiv preprint arXiv:2512.22336*.

Jie Huang and Kevin Chang. 2023. Large language models as general pattern machines. *ArXiv preprint*, abs/2306.09831.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, Yikai Zhang, Yuqing Yang, Ting Wu, Binjie Wang, Shichao Sun, Yang Xiao, Yiyuan Li, Fan Zhou, Steffi Chern, and 9 others. 2024. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent AI. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Zhiyuan Huang, Baichuan Yang, Zikun He, Yanhong Wu, Fang Hongyu, Zhenhe Liu, Lin Dongsheng, and Bing Su. 2025b. Chemvts-bench: Evaluating visual-textual-symbolic reasoning of multimodal large language models in chemistry. *ArXiv preprint*, abs/2511.17909.

Hyeon Hwang, Yewon Cho, Chanwoong Yoon, Yein Park, Minju Song, Kyungjae Lee, Gangwoo Kim, and Jaewoo Kang. 2025. Assessing llm reasoning steps via principal knowledge grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19925–19948.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. 2024. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. *ArXiv preprint*, abs/2402.00559.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *ArXiv preprint*, abs/2403.07974.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Transfer learning for natural language processing workshop*, pages 52–62. PMLR.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023b. Towards mitigating hallucination in large language models via self-reflection. *arXiv preprint arXiv:2310.06271*.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, and 1 others. 2025a. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *ArXiv preprint*, abs/2502.09621.

Gangwei Jiang, Yahui Liu, Zhaoyi Li, Wei Bi, Fuzheng Zhang, Linqi Song, Ying Wei, and Defu Lian. 2025b. What makes a good reasoning chain? uncovering structural patterns in long chain-of-thought reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6501–6525.

Yuan Jiang, Yujian Zhang, Liang Lu, Christoph Treude, Xiaohong Su, Shan Huang, and Tiantian Wang. 2025c. Enhancing high-quality code generation in large language models with comparative prefix-tuning. *ArXiv preprint*, abs/2503.09020.

Zheng Ping Jiang, Yining Lu, Hanjie Chen, Daniel Khashabi, Benjamin Van Durme, and Anqi Liu. 2024a. Rora: Robust free-text rationale evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1070–1087.

Zhuoxuan Jiang, Haoyuan Peng, Shanshan Feng, Fan Li, and Dongsheng Li. 2024b. Llms can find mathematical reasoning mistakes by pedagogical chain-of-thought. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 3439–3447. ijcai.org.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step length on large language models. *ArXiv preprint*, abs/2401.04925.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. GRACE: Discriminator-guided chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15299–15328, Singapore. Association for Computational Linguistics.

Heegyu Kim, Taeyang Jeon, Seunghwan Choi, Seungtaek Choi, and Hyunsouk Cho. 2024a. Flex: Expert-level false-less execution metric for reliable text-to-sql benchmark. *ArXiv preprint*, abs/2409.19014.

Seungone Kim, Jamin Shin, Yejin Choi, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024b. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Aakriti Kumar, Nalin Poungpeth, Diyi Yang, Erina Farrell, Bruce Lambert, and Matthew Groh. 2025. When large language models are reliable for judging empathic communication. *ArXiv preprint*, abs/2506.10150.

Lucas Fonseca Lage and Simon Ostermann. 2025. Openfactscore: Open-source atomic evaluation of factuality in text generation. *ArXiv preprint*, abs/2507.05965.

Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Wen-Tau Yih, Daniel Fried, Sida I. Wang, and Tao Yu. 2023. DS-1000: A natural and reliable benchmark for data science code generation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 18319–18345. PMLR.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, and 1 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *ArXiv preprint*, abs/2307.13702.

Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. 2024. Lab-bench: Measuring capabilities of language models for biology research. *ArXiv preprint*, abs/2407.10362.

Joshua Ong Jun Leang, Aryo Pradipta Gema, and Shay B Cohen. 2025. Comat: Chain of mathematically annotated thought improves mathematical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20256–20285.

Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. 2024. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100213.

Jaehyeok Lee, Keisuke Sakaguchi, and JinYeong Bak. 2025a. Self-training meets consistency: Improving llms' reasoning with consistency-driven rationale evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10519–10539.

Jinu Lee and Julia Hockenmaier. 2025. Evaluating step-by-step reasoning traces: A survey. *ArXiv preprint*, abs/2502.12289.

Young-Jun Lee, Seungone Kim, Byung-Kwan Lee, Minkyeong Moon, Yechan Hwang, Jong Myoung Kim, Graham Neubig, Sean Welleck, and Ho-Jin Choi. 2025b. Refinebench: Evaluating refinement capability of language models via checklists. *arXiv preprint arXiv:2511.22173*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay V. Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Bohan Li, Jiannan Guan, Longxu Dou, Yunlong Feng, Dingzirui Wang, Yang Xu, Enbo Wang, Qiguang Chen, Bichen Wang, Xiao Xu, and 1 others. 2025a. Can large language models understand you better? an mbti personality detection dataset aligned with population traits. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5071–5081.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025b. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Jia Li, Ge Li, Yongmin Li, and Zhi Jin. 2025c. Structured chain-of-thought prompting for code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2):1–23.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024a. More agents is all you need. *arXiv preprint arXiv:2402.05120*.

Ruosen Li and Xinya Du. 2023. Leveraging structured information for explainable multi-hop question answering and reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6779–6789, Singapore. Association for Computational Linguistics.

Ruosen Li, Ziming Luo, and Xinya Du. 2024b. Fine-grained hallucination detection and mitigation in language model mathematical reasoning.

Xiaoyuan Li, Wenjie Wang, Moxin Li, Junrong Guo, Yang Zhang, and Fuli Feng. 2024c. Evaluating mathematical reasoning of large language models: A focus on error identification and correction. *ArXiv preprint*, abs/2406.00755.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023b. Diverse: A reliable approach to few-shot reasoning. In *ACL*.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. 2024d. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, and 1 others. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *ArXiv preprint*, abs/2211.09110.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *EMNLP*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023a. Deductive verification of chain-of-thought reasoning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhan Ling and 1 others. 2023b. Thought propagation: An analogical approach to complex reasoning. volume abs/2310.03965.

Chengwu Liu, Ye Yuan, Yichun Yin, Yan Xu, Xin Xu, Zaoyu Chen, Yasheng Wang, Lifeng Shang, Qun Liu, and Ming Zhang. 2025a. Safe: Enhancing mathematical reasoning in large language models via retrospective step-aware formal verification. *ArXiv preprint*, abs/2506.04592.

Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. 2023a. LogiCoT: Logical chain-of-thought instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2908–2921, Singapore. Association for Computational Linguistics.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Tianqiao Liu, Zui Chen, Zitao Liu, Mi Tian, and Weiqi Luo. 2024. Expediting and elevating large language model reasoning via hidden chain-of-thought decoding. *ArXiv preprint*, abs/2409.08561.

Tongxuan Liu, Wenjiang Xu, Weizhe Huang, Yuting Zeng, Jiaxing Wang, Xingyu Wang, Hailong Yang, and Jing Li. 2025b. Logic-of-thought: Injecting logic into contexts for full reasoning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10168–10185.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, and 1 others. 2021. Codexglue: A machine learning benchmark

dataset for code understanding and generation. *ArXiv preprint*, abs/2102.04664.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2024. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, D Deligiannis, and 1 others. 2024. Walk the talk? measuring the faithfulness of large language model explanations. In *ICLR*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023a. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023b. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.

Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. 2023a. What makes chain-of-thought prompting effective? a counterfactual study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1448–1535, Singapore. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023b. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2024. Discoverybench: Towards data-driven discovery with large language models. *ArXiv preprint*, abs/2407.01725.

Youssef Maklad, Fares Wael, Wael Elsersy, and Ali Hamdi. 2025. Retrieval augmented generation based llm evaluation for protocol state machine inference with chain-of-thought reasoning. In *International Congress on Information and Communication Technology*, pages 313–323. Springer.

Zhenjiang Mao, Artem Bisliouk, Rohith Reddy Nama, and Ivan Ruchkin. 2025. Temporalizing confidence: Evaluation of chain-of-thought reasoning with signal temporal logic. *ArXiv preprint*, abs/2506.08243.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2024. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stephen Miner, Yoshiki Takashima, Simeng Han, Sam Kouteili, Ferhat Erata, Ruzica Piskac, and Scott J Shapiro. 2024. Scheherazade: Evaluating chain-of-thought math reasoning in llms with chain-of-problems. *ArXiv preprint*, abs/2410.00151.

Francesco Maria Molfese, Luca Moroni, Ciro Porcaro, Simone Conia, and Roberto Navigli. 2025. Retraceqa: Evaluating reasoning traces of small language models in commonsense question answering. *ArXiv preprint*, abs/2510.09351.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on*

Empirical Methods in Natural Language Processing, pages 20286–20332.

S Mysore and 1 others. 2023. Crisp: Complex reasoning with interpretable step-based plans. In *EMNLP*.

Sina Bagheri Nezhad, Yao Li, and Ameeta Agrawal. 2025. Symcode: A neurosymbolic approach to mathematical reasoning via verifiable code generation. *ArXiv preprint*, abs/2510.25975.

Minh-Vuong Nguyen, Linhao Luo, Fatemeh Shiri, Dinh Phung, Yuan-Fang Li, Thuy Vu, and Gholamreza Haffari. 2024. Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2862–2883.

Phuong Minh Nguyen, Tien Huu Dang, and Naoya Inoue. 2025. Non-interactive symbolic-aided chain-of-thought for logical reasoning. *ArXiv preprint*, abs/2508.12425.

Natapong Nitarach, Warit Sirichotedumrong, Panop Pitchayarthorn, Pittawat Taveekitworachai, Potsawee Manakul, and Kunat Pipatanakul. 2025. Fincot: Grounding chain-of-thought in expert financial reasoning. *ArXiv preprint*, abs/2506.16123.

Maxwell Nye and 1 others. 2021. Show your work: Scratchpads for intermediate computation with language models. *ArXiv preprint*, abs/2112.00114.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, Singapore. Association for Computational Linguistics.

OpenAI. 2025. Gpt-5 system card. Technical report, OpenAI. Accessed: 2025-08-10.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. Memgpt: Towards llms as operating systems.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore. Association for Computational Linguistics.

Atharva Pandey, Kshitij Dubey, Rahul Sharma, and Amit Sharma. 2025. Deduce: Deductive consistency as a framework to evaluate llm reasoning. *ArXiv preprint*, abs/2504.07080.

Shrey Pandit, Austin Xu, Xuan-Phi Nguyen, Yifei Ming, Caiming Xiong, and Shafiq Joty. 2025. Hard2verify: A step-level verification benchmark for open-ended frontier math. *ArXiv preprint*, abs/2510.13744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Kaviraj Pather, Elena Hadjigeorgiou, Arben Krasniqi, Claire Schmit, Irina Rusu, Marc Pons, and Kabir Khan. 2025. Vis-cot: A human-in-the-loop framework for interactive visualization and intervention in llm chain-of-thought reasoning. *ArXiv preprint*, abs/2509.01412.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. *ArXiv preprint*, abs/2402.13950.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and 1 others. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *ArXiv preprint*, abs/2302.12813.

Gabriel Poesia, Kanishk Gandhi, Eric Zelikman, and Noah D Goodman. 2023. Certified deductive reasoning with language models. *ArXiv preprint*, abs/2306.04031.

Nearchos Potamitis, Lars Klein, and Akhil Arora. 2025. Reasonbench: Benchmarking the (in) stability of llm reasoning. *arXiv preprint arXiv:2512.07795*.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. ReCEval: Evaluating reasoning chains via correctness and informativeness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10066–10086, Singapore. Association for Computational Linguistics.

Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. 2025. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. *ArXiv preprint*, abs/2504.07956.

Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, and 1 others. 2024. Scaling large language model-based multi-agent collaboration. *arXiv preprint arXiv:2406.07155*.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang,

and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *arXiv preprint arXiv:2310.14799*.

Libo Qin, Qiguang Chen, Jingxuan Zhou, Jin Wang, Hao Fei, Wanxiang Che, and Min Li. 2025. Divide-solve-combine: An interpretable and accurate prompting framework for zero-shot multi-intent detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25038–25046.

Jack W Rae and 1 others. 2021. Scaling language models: Methods, analysis and insights from training gopher. *ArXiv preprint*, abs/2112.11446.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sahana Ramnath, Brihi Joshi, Skyler Hallinan, Ximing Lu, Liunian Harold Li, Aaron Chan, Jack Hessel, Yejin Choi, and Xiang Ren. 2024. Tailoring self-rationalizers with multi-reward distillation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.

Jon Saad-Falcon, Rajan Vivek, William Berrios, Nandita Shankar Naik, Matija Franklin, Bertie Vidgen, Amanpreet Singh, Douwe Kiela, and Shikib Mehri. 2024. Lmunit: Fine-grained evaluation with natural language unit tests. *ArXiv preprint*, abs/2412.13091.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2024. Branch-solve-merge improves large language model evaluation and generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8352–8370, Mexico City, Mexico. Association for Computational Linguistics.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Hekar A Mohammed Salih and Qusay I Sarhan. 2025. A study of large language models in detecting python code violations. *ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY*, 13(2):215–225.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2023. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4454–4470, Toronto, Canada. Association for Computational Linguistics.

Jianshu She, Zhuohao Li, Zhemin Huang, Qi Li, Peiran Xu, Haonan Li, and Qirong Ho. 2025. Hawkeye:

Model collaboration for efficient reasoning. In *Second Conference on Language Modeling*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Toby Simonds, Kevin Lopez, Akira Yoshiyama, and Dominique Garmier. 2025. Rlsr: Reinforcement learning from self reward.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025. Prmbench: A fine-grained and challenging benchmark for process-level reward models. *arXiv preprint arXiv:2501.03124*.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

Andreas Stephan, Dawei Zhu, Matthias Aßenmacher, Xiaoyu Shen, and Benjamin Roth. 2025. From calculation to adjudication: Examining llm judges on mathematical reasoning tasks. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 759–773.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *ArXiv preprint*, abs/2405.01724.

Alane Suhr, Clara Vania, Nikita Nangia, Maarten Sap, Mark Yatskar, Samuel R. Bowman, and Yoav Artzi. 2021. Crowdsourcing beyond annotation: Case studies in benchmark data collection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 1–6, Punta Cana, Dominican Republic & Online. Association for Computational Linguistics.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, and 1 others. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19053–19061. AAAI Press.

Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 952–966.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Alon Talmor and 1 others. 2023. Evaluating the reasoning abilities of large language models. *ArXiv preprint*, abs/2310.06828.

Megh Thakkar, Tolga Bolukbasi, Sriram Ganapathy, Shikhar Vashishth, Sarath Chandar, and Partha Talukdar. 2023. Self-influence guided data reweighting for language model pre-training. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2033–2045, Singapore. Association for Computational Linguistics.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, Hao Tong, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Shengzhu Yin, and 10 others. 2024. Scicode: A research coding benchmark curated by scientists. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. 2025. Measuring faithfulness of chains of thought by unlearning reasoning steps. *arXiv e-prints*, pages arXiv–2502.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *ArXiv preprint*, abs/2211.14275.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Hemish Veeraboina. 2023. Aime problem set 1983-2024.

Theo Walker, Christopher M Grulke, Diane Pozefsky, and Alexander Tropsha. 2010. Chembench: a cheminformatics workbench. *Bioinformatics*, 26(23):3000–3001.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Boxuan Wang, Zhuoyun Li, Xinmiao Huang, Xiaowei Huang, and Yi Dong. 2025a. Chasing consistency: Quantifying and optimizing human-model alignment in chain-of-thought reasoning. *ArXiv preprint*, abs/2511.06168.

Changyue Wang, Weihang Su, Qingyao Ai, and Yiqun Liu. 2025b. Joint evaluation of answer and reasoning consistency for hallucination detection in large reasoning models. *ArXiv preprint*, abs/2506.04832.

Jianing Wang, Qiushi Sun, Xiang Li, and Ming Gao. 2024a. Boosting language models reasoning with chain-of-knowledge prompting. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4958–4981.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024b. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Linhao Wang, Zihan Wang, Jinghao Lin, Qiwei Zeng, Jingdan Zhang, Tianrun Wu, Xuanhao Zhang, Junjie Li, Zijing Wang, Yongkang Liu, and 1 others. 2025c. Good teachers, better students: A survey of reward models for llm. *Authorea Preprints*.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024c. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439.

Tianyu Wang, Nianjun Zhou, and Zhixiong Chen. 2024d. Enhancing computer programming education with llms: A study on effective prompt engineering for python code generation. *ArXiv preprint*, abs/2407.05437.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024e. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2024f. MINT: evaluating llms in multi-turn interaction with tools and language feedback. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Xinyuan Wang and 1 others. 2023b. Is reasoning a hindsight? evaluating the faithfulness of chain-of-thought explanations. In *EMNLP*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023d. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024g. Pandalm: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yu Wang, Shiwan Zhao, Zhihu Wang, Heyuan Huang, Ming Fan, Yubo Zhang, Zhixing Wang, Haijun Wang, and Ting Liu. 2024h. Strategic chain-of-thought: Guiding accurate reasoning in llms through strategy elicitation. *ArXiv preprint*, abs/2409.03271.

Zijian Wang and Chang Xu. 2025. Thoughtprobe: Classifier-guided llm thought space exploration via probing representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6029–6050.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575.

Junda Wu, Xintong Li, Ruoyu Wang, Yu Xia, Yuxin Xiong, Jianing Wang, Tong Yu, Xiang Chen, Branislav Kveton, Lina Yao, and 1 others. 2024a. Ocean: Offline chain-of-thought evaluation and alignment in large language models. *ArXiv preprint*, abs/2410.23703.

Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan Rossi, Sungchul Kim, Anup Rao, and Julian McAuley. 2024b. Decot: Debiasing chain-of-thought for knowledge-intensive tasks in large language models via causal intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14073–14087.

Xueqing Wu, Yuheng Ding, Bingxuan Li, Pan Lu, Da Yin, Kai-Wei Chang, and Nanyun Peng. 2025a. Visco: Benchmarking fine-grained critique and correction towards self-improvement in visual reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9527–9537.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2025b. Inference scaling laws: An empirical analysis of compute-optimal inference for llm problem-solving. In *The Thirteenth International Conference on Learning Representations*.

Yuyang Wu, Yifei Wang, Ziyu Ye, Tianqi Du, Stefanie Jegelka, and Yisen Wang. 2025c. When more is less: Understanding chain-of-thought length in llms. *ArXiv preprint*, abs/2502.07266.

Yu Xia, Rui Wang, Xu Liu, Mingyan Li, Tong Yu, Xiang Chen, Julian McAuley, and Shuai Li. 2025. Beyond chain-of-thought: A survey of chain-of-x paradigms for llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10795–10809.

Jiaqi Xu, Cuiling Lan, Xuejin Chen, and Yan Lu. 2025. Stepwise think-critique: A unified framework for robust and interpretable llm reasoning. *arXiv preprint arXiv:2512.15662*.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024a. Faithful logical reasoning via symbolic chain-of-thought. *ArXiv preprint*, abs/2405.18357.

Xin Xu, Shizhe Diao, Can Yang, and Yang Wang. 2024b. Can we verify step by step for incorrect answer detection? *ArXiv preprint*, abs/2402.10528.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *ArXiv preprint*, abs/2505.09388.

Sohee Yang, Sang-Woo Lee, Nora Kassner, Daniela Gottesman, Sebastian Riedel, and Mor Geva. 2025b. How well can reasoning models identify and recover from unhelpful thoughts? *ArXiv preprint*, abs/2506.10979.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. 2023a. Satlm: Satisfiability-aided language models using declarative prompting. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Xi Ye and 1 others. 2023b. A comprehensive study of chain-of-thought reasoning in large language models: Advances, challenges, and opportunities. *ArXiv preprint*, abs/2309.15402.

Wenhan Yu, Xinbo Lin, Lanxin Ni, Jinhua Cheng, and Lei Sha. 2025. Benchmarking multi-step legal reasoning and analyzing chain-of-thought effects in large language models. *ArXiv preprint*, abs/2511.07979.

Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023. Towards better chain-of-thought prompting strategies: A survey. *ArXiv preprint*, abs/2310.04959.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Weiqi Yue, Yuyu Yin, Xin Zhang, Binbin Shi, Tingting Liang, and Jian Wan. 2025. Cot4rec: Revealing user preferences through chain of thought for recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 13142–13151.

Zhongshen Zeng, Pengguang Chen, Shu Liu, Haiyun Jiang, and Jiaya Jia. 2023. Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation. *ArXiv preprint*, abs/2312.17080.

Bingyuan Zhang, Xulong Zhang, Yong Zhang, Jun Yu, and Jianzong Wang. 2025a. Logic consistency makes large language models personalized reasoning teachers. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Boning Zhang, Chengxi Li, and Kai Fan. 2024a. Mario eval: Evaluate your math llm with your math llm–a mathematical dataset evaluation toolkit. *ArXiv preprint*, abs/2404.13925.

Jieyu Zhang, Ranjay Krishna, Ahmed H Awadallah, and Chi Wang. 2023a. Ecoassistant: Using llm assistant more affordably and accurately. *ArXiv preprint*, abs/2310.03046.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2025b. Generative verifiers: Reward modeling as next-token prediction. In *The Thirteenth International Conference on Learning Representations*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024b. How language model hallucinations can snowball. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, and 1 others. 2025c. A survey on test-time scaling in large language models: What, how, where, and how well? *arXiv preprint arXiv:2503.24235*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. 2024c. Chain of preference optimization: Improving chain-of-thought reasoning in llms. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025d. The lessons of developing process reward models in mathematical reasoning. *ArXiv preprint*, abs/2501.07301.

Zhuosheng Zhang, Yao Yao, Aston Zhang, Xiangru Tang, Xinbei Ma, Zhiwei He, Yiming Wang, Mark Gerstein, Rui Wang, Gongshen Liu, and 1 others. 2025e. Igniting language intelligence: The hitchhiker's guide from chain-of-thought reasoning to language agents. *ACM Computing Surveys*, 57(8):1–39.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

Kesen Zhao, Beier Zhu, Qianru Sun, and Hanwang Zhang. 2025a. Unsupervised visual chain-of-thought reasoning via preference optimization. *arXiv preprint arXiv:2504.18397*.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5823–5840, Toronto, Canada. Association for Computational Linguistics.

Zheng Zhao, Yeskendir Koishekenov, Xianjun Yang, Naila Murray, and Nicola Cancedda. 2025b. Verifying chain-of-thought reasoning via its computational graph. *ArXiv preprint*, abs/2510.09312.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Runtao Zhou, Giang Nguyen, Nikita Kharya, Anh Nguyen, and Chirag Agarwal. 2025. Improving human verification of llm reasoning through interactive explanation interfaces. *ArXiv preprint*, abs/2510.22922.

Zhanke Zhou, Rong Tao, Jianing Zhu, Yiwen Luo, Zengmao Wang, and Bo Han. 2024. Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales? In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Kaijie Zhu, Qinlin Zhao, Hao Chen, Jindong Wang, and Xing Xie. 2024. Promptbench: A unified library for evaluation of large language models. *Journal of Machine Learning Research*, 25(254):1–22.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, and 1 others. 2023. Mindstorms in natural language-based societies of mind. *arXiv preprint arXiv:2305.17066*.

A Example of Disguised Accuracy

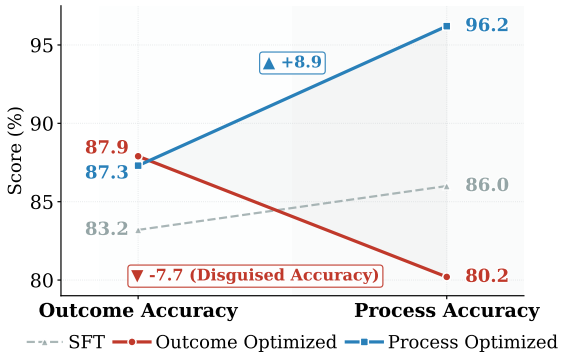


Figure 5: An example showing how high answer accuracy can mask poor intermediate reasoning.

Figure 5 illustrates experimental results from (Uesato et al., 2022) using a 70B model trained with Supervised Fine-Tuning (SFT), Outcome-Optimized, and Process-Optimized methods, showing the disguised accuracy phenomenon (Bentham et al., 2024). In these experiments, both the answer accuracy and the accuracy of the reasoning process are evaluated on the same dataset. The results show that while the Outcome-Optimized method improves final outcome accuracy, it causes a severe degradation in the quality of intermediate reasoning steps. This indicates that the model guesses correct answers using flawed logic rather than sound reasoning. As a result, this phenomenon leads to a reversal in model rankings, where a model with higher outcome metrics actually demonstrates inferior reasoning capabilities. In practical applications, such models provide low trustworthiness because their high accuracy masks fundamental reasoning failures.

B Related Work

Surveys on CoT Generation and Prompting. A substantial body of literature reviews the landscape of CoT reasoning, primarily focusing on generation strategies and prompt engineering (Yu et al., 2023; Chu et al., 2024; Zhang et al., 2025e; Qiao et al., 2023). These surveys extensively catalog paradigms ranging from Zero-shot (Kojima et al., 2022) and Few-shot CoT (Wei et al., 2022) to complex agentic frameworks (Chen et al., 2025c; Xia et al., 2025). However, their analytical lens is predominantly cast on *how to elicit* reasoning rather than *how to evaluate* it. Consequently, assessment in these works is often reduced to outcome-oriented metrics (e.g., accuracy on GSM8K), largely conflating the validity of the intermediate process with the correctness of the final answer (Lee and Hockenmaier, 2025).

General LLM Evaluation and Specific Subdomains. Broader reviews of LLM evaluation (Chang et al., 2024; Guo et al., 2023; Liang et al., 2022; Srivastava et al., 2023) treat reasoning as merely one of many capabilities. While they establish benchmarks for general performance, they rarely dissect the granular quality of reasoning traces, such as redundancy or step-wise necessity. Similarly, while surveys on LLM-as-a-Judge (Wang et al., 2024b; Zheng et al., 2023) and hallucination (Ji et al., 2023a; Huang et al., 2025a) touch upon evaluation mechanics, they typically focus on linguistic fluency, preference alignment, or factual fabrication, lacking a systematic taxonomy for the causal rigor and logical consistency inherent to multi-step reasoning.

To date, there is no comprehensive survey dedicated specifically to CoT evaluation that systematically bridges theoretical dimensions with practical methodologies across diverse domains. Our work addresses this critical void by establishing a unified framework structured along four core dimensions: ***What*** to measure, ***How*** to verify, ***Who*** evaluates, and ***Where*** to apply.

C Itegration of Interpretability Methods

To assess CoT fidelity, recent studies integrate the explainability methods. Researchers utilize tools like LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), alongside gradient-based self-influence methods (Thakkar et al., 2023), to cross-verify if implicit token importance aligns with explicit reasoning steps. Counterfactual perturbations test causal validity by determining whether the generated rationale consistently dictates the prediction shift under input changes (Turpin et al., 2023; Atanasova et al., 2023; Wang et al., 2023b). This combined framework helps distinguish authentic reasoning processes from persuasive yet unfaithful post-hoc rationalizations (Lyu et al., 2024; Lanham et al., 2023).

D Benchmarks & Evaluation Methodologies

Table 1 provides an overview of the benchmark datasets across different tasks mentioned in this survey for CoT evaluation.

D.1 Mathematical Reasoning

D.1.1 Benchmarks

Mathematical evaluation prioritizes absolute **Correctness**, yet complex problems demand higher **Logical Coherence**. Basic robustness relies on GSM8K (Cobbe et al., 2021) and SVAMP (Patel et al., 2021). Conversely, MathQA (Amini et al., 2019) and MATH (Hendrycks et al., 2021) target interpretable skills where **Cognitive Efficiency** often yields to detailed derivation. MATH further includes challenging problems like AMC12 and AIME (Veeraboina, 2023). To address saturation, recent works increase complexity to test **Logical Coherence**. GSM8K-Scheherazade (Miner et al., 2024) uses chained queries, while CoMAT (Leang et al., 2025) and SCoT (Wang et al., 2024h) enforce strategy elicitation.

D.1.2 Evaluation Methodologies

Automated Verification. Automated strategies split into symbolic verification for **Correctness** and semantic scoring for **Interpretability**. Tools like MathSensei (Das et al., 2024), SymCode (Nezhad et al., 2025), and RLSR (Simonds et al., 2025) leverage SymPy for execution, while Safe (Liu et al., 2025a) uses retrospective verification. To assess **Logical Coherence**, verifiers from GSM8K (Cobbe et al., 2021) evolve into step-wise discriminators in CoMAT (Leang et al., 2025) and Hidden CoT (Liu et al., 2024). Additionally, Nguyen et al. (2024) use knowledge graphs for semantic alignment, aggregated by MARIO Eval (Zhang et al., 2024a).

Human-in-the-Loop. Human evaluation remains the ground truth. Experts are essential for assessing **Logical Coherence** in complex proofs (Zeng et al., 2023) and providing process supervision (Lightman et al., 2024), serving as a gold standard (Nguyen et al., 2024). Conversely, crowdsourcing offers scalability for subjective **Interpretability** in ROSCOE (Golovneva et al., 2023), despite limitations in complex logic. Interactive interfaces (Zhou et al., 2025) help mitigate this by visualizing reasoning chains for non-experts.

D.2 Commonsense Reasoning

D.2.1 Benchmarks

Commonsense domains prioritize **Interpretability** and **Cognitive Efficiency**. Benchmarks include CommonsenseQA (Talmor et al., 2019), StrategyQA (Geva et al., 2021), ARC (Clark et al., 2018), and OpenBookQA (Mihaylov et al., 2018). ROSCOE (Golovneva et al., 2023) quantifies semantic **Consistency**, while SELF-CHECK (Miao et al., 2024) detects errors. To ensure **Correctness**, SciKnowEval (Feng et al., 2024) and CoK (Wang et al., 2024a) integrate knowledge triples, while De-CoT (Wu et al., 2024b) mitigates bias. VISCO (Wu et al., 2025a) extends this to multimodal tasks.

D.2.2 Evaluation Methodologies

Hybrid Evaluation. Hybrid methods are common, using frameworks like RORA (Jiang et al., 2024a) and ACORN (Brassard et al., 2024) to align rationales with human **Interpretability**. Along with ReTraceQA (Molfese et al., 2025), these show that open-weight models often need specific prompting to match human intuition (Zheng et al., 2023; Bang et al., 2023).

Analysis of Pitfall. Methodologies must diagnose violations of **Cognitive Efficiency**, such as bias or over-reasoning. CoT can amplify stereotypes (Shaikh et al., 2023), while over-explaining axioms leads to hallucinations. Protocols test for unfaithful explanations that fail to support predictions (Turpin et al., 2023).

D.3 Code Reasoning

D.3.1 Benchmarks

Code reasoning demands strict **Correctness** via execution while using **Logical Coherence** for planning. Benchmarks range from HumanEval (Chen, 2021) and MBPP (Austin et al., 2021) to competition sets like APPS (Hendrycks et al., 2021) and CodeContests (Li et al., 2022). Multilingual tests include MBXP and MathQA-X (Athiwaratkun et al., 2023). Specialized domains cover data science in DS-1000 (Lai et al., 2023) and general understanding in CodeXGLUE (Lu et al., 2021). Recent tools like CRUXEval (Gu et al., 2024) and LiveCodeBench (Jain et al., 2024) expand this to execution reasoning.

D.3.2 Evaluation Methodologies

Automated Evaluation. Protocols assess functional correctness via execution metrics like pass@k (Chen, 2021). Complementing this, static analysis tools like Pylint detect syntactic violations and enforce style **Consistency**, identifying latent errors without execution (Salih and Sarhan, 2025; Jiang et al., 2025c; Almeida et al., 2024; Wang et al., 2024d; Al-Khafaji and Majeed, 2024).

Human Evaluation. Human review verifies algorithmic rationality and **Logical Coherence**. Experts audit code for theoretical soundness, examining subtle edge cases that automated suites might miss (Dunivin, 2024; Li et al., 2025c, 2022; Austin et al., 2021).

D.4 Dialogue & Multi-turn Reasoning

D.4.1 Benchmarks

In multi-turn dialogue, **Consistency** is dominant. MT-Bench (Zheng et al., 2023) and MINT (Wang et al., 2024f) assess stability. MuTual (Cui et al., 2020) and DREAM (Sun et al., 2019) focus on logic. QReCC (Anantha et al., 2021) and CICERO (Ghosal et al., 2022) target state tracking and causal inference, while MUSR (Sprague et al., 2024) and ClemBench (Chalamalasetti et al., 2023) test long-context rules.

D.4.2 Evaluation Methodologies

Automated Evaluation. Protocols decompose structure to verify **Logical Coherence**. The Q^2 framework (Honovich et al., 2021) generates constituent questions for factual support. DNLI (Welleck et al., 2019) automates **Consistency** checks by classifying logical relationships between utterances.

Human Evaluation. Human evaluation audits cross-turn **Consistency**. Experts ensure coherent belief states without violating constraints. Zheng et al. (2023) formalize this in MT-Bench using pairwise comparisons to penalize instruction drift (Suhr et al., 2021).

D.5 Scientific Reasoning

Scientific evaluation places a premium on **Correctness** and **Logical Coherence**. GPQA (Rein et al., 2024) sets a PhD-level standard, while MegaScience (Fan et al., 2025) addresses data scarcity. SCIBench (Wang et al., 2024e) and OlympicArena (Huang et al., 2024) test cognitive capabilities. Specialized benchmarks include ChemVTS-Bench (Huang et al., 2025b) and ChemBench (Walker et al., 2010) for chemistry, LAB-Bench (Laurent et al., 2024) for biology, and SciCode (Tian et al., 2024) for physics. Multimodal frontiers like ScienceQA (Saikh et al., 2022) and SciEval (Sun et al., 2024) lead to Discovery-Bench (Majumder et al., 2024), challenging agents to formulate hypotheses.

E Extended Discussion on Future Directions

This appendix provides a detailed analysis of the four strategic frontiers for CoT evaluation outlined in Section 6.

E.1 Hybrid Evaluation Paradigms and Standardization

The dichotomy between expensive human annotation and potentially biased automated evaluation remains a bottleneck. Future work should pivot towards hybrid human-AI paradigms, where models perform initial filtering or semi-automated scoring to reduce workload, followed by targeted human verification of uncertain cases (Diao et al., 2024; Li et al., 2023a). This approach mitigates the high cost and subjectivity of manual review while maintaining reliability. Furthermore, to address the lack of standardization across studies, the community must establish unified annotation guidelines and consistent evaluation protocols. Research into **adversarial robustness** is also vital, employing techniques to generate adversarial samples that stress-test the stability of reasoning evaluation metrics against subtle prompt variations (Cui et al., 2025; Sclar et al., 2024; Zhou et al., 2024).

E.2 Multimodal and Multilingual Frontiers

As models expand beyond text, evaluation frameworks must adapt to **Multimodal CoT**. This involves developing MLLM-as-a-Judge frameworks capable of assessing reasoning that intertwines visual perception with textual logic, ensuring reliability under distribution shifts (Jiang et al., 2025a; Chen et al., 2024b; Wu et al., 2025a). Simultaneously, the Anglocentric bias in current benchmarks necessitates a shift towards multilingual evaluation, particularly for low-resource and minority languages. Initiatives like the Aya project and multilingual benchmarks (Bang et al., 2023; Athiwaratkun et al., 2023) demonstrate the importance of instruction-tuned evaluation across diverse linguistic landscapes to ensure equitable reasoning capabilities.

E.3 Quantifying Interpretability and Trustworthiness

Moving beyond vague notions of explainability, future research must mathematically quantify the structural validity of reasoning. Promising directions include leveraging Causal Graphs and Knowl-

edge Graphs to map the logical flow of CoT, transforming text-based traces into verifiable structural paths (Amayuelas et al., 2025; Nguyen et al., 2024; Zhao et al., 2025b; Luo et al., 2024). Additionally, integrating Explainable AI techniques can facilitate human-AI collaborative review, allowing evaluators to pinpoint exactly which part of the input context influenced a specific reasoning step, thereby addressing the challenge of disguised accuracy where models arrive at correct answers via flawed logic (Thakkar et al., 2023; Wang and Xu, 2025; Turpin et al., 2023).

CoT evaluation should also transition from general benchmarks to high-stakes vertical applications. In domains such as healthcare, law, and education, evaluation metrics must prioritize trustworthiness, evidence retrieval, and safety over simple accuracy. For instance, legal reasoning requires validating adherence to statutory interpretations (Yu et al., 2025), while educational applications demand iterative self-correction and feedback mechanisms for pedagogical effectiveness (Jiang et al., 2024b; Wang et al., 2024d). Moreover, bias analysis is crucial; frameworks must be developed to detect deep-seated stereotypes (e.g., gender or cultural bias) within reasoning chains (Shaikh et al., 2023; Wu et al., 2024b). In enterprise contexts, evaluation protocols must also align with auditability and compliance standards to support safe large-scale deployment.

E.4 Probing Reasoning Upper Bounds

Current benchmarks largely measure static inference, overlooking the potential of test-time scaling (TTS) where models simulate deliberate thinking (Snell et al., 2024; Goyal et al., 2023). Future evaluation must quantify the efficiency-performance trade-off, verifying if extended reasoning time and search depth yield proportional accuracy gains (Brown et al., 2024; Wu et al., 2025b). This shift necessitates new benchmarks like JETTS to ensure automated judges can reliably monitor these complex processes (Lightman et al., 2024; Uesato et al., 2022). Simultaneously, protocols must expand to multi-agent systems, assessing whether increasing ensemble size follows predictable scaling laws or encounters diminishing returns due to redundancy (Li et al., 2024a; Zhuge et al., 2023).

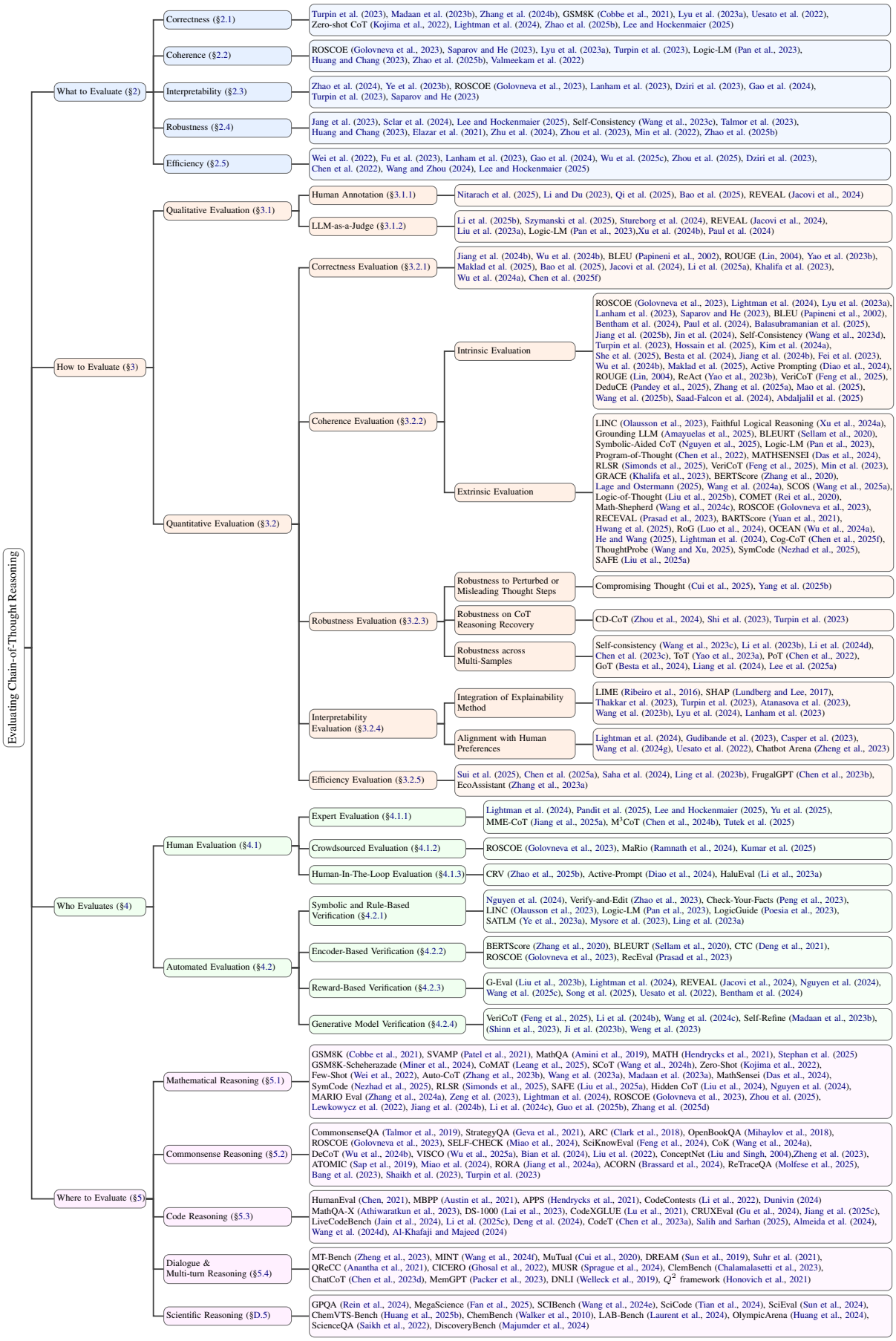


Figure 6: Our full taxonomy for the survey on Evaluating Chain-of-Thought Reasoning.

Table 1: Overview of benchmarks used in CoT evaluation across different tasks.

Task	Dataset	Description
Mathematical Reasoning	**GSM8K** (Cobbe et al., 2021)	8.5K grade school math word problems with step-by-step solutions.
	SVAMP (Patel et al., 2021)	Challenge set testing robustness to superficial keywords.
	MathQA (Amini et al., 2019)	GRE-level math problems with annotated operation programs.
	MATH (Hendrycks et al., 2021)	12,500 competition math problems (AMC10/12, AIME).
	AIME (Veeraboina, 2023)	High-difficulty competition problems (often included/covered by MATH).
	GSM8K-Scheherazade (Miner et al., 2024)	Synthetic extension of GSM8K to counter saturation.
Commonsense Reasoning	**CommonsenseQA** (Talmor et al., 2019)	Multiple-choice questions requiring commonsense knowledge.
	StrategyQA (Geva et al., 2021)	Implicit reasoning steps for multi-hop reasoning.
	ARC (Clark et al., 2018)	Science exam questions testing genuine understanding.
	OpenBookQA (Mihaylov et al., 2018)	Questions requiring knowledge retrieval and reasoning.
	ROSCOE (Golovneva et al., 2023)	Metrics for semantic consistency in reasoning chains.
	SELF-CHECK (Miao et al., 2024)	Automated step-wise error detection framework.
	SciKnowEval (Feng et al., 2024)	Scientific knowledge evaluation with knowledge triples.
	CoK (Wang et al., 2024a)	Chain-of-Knowledge framework for evidence retrieval.
	DeCoT (Wu et al., 2024b)	Causal intervention framework for bias mitigation.
	VISCO (Wu et al., 2025a)	Multimodal visual reasoning evaluation.
Code Reasoning	**HumanEval** (Chen, 2021)	164 hand-written programming problems with pass@k metrics.
	MBPP (Austin et al., 2021)	Basic Python problems dataset.
	APPS (Hendrycks et al., 2021)	Competition-level programming problems.
	CodeContests (Li et al., 2022)	Programming competition problems from various platforms.
	MBXP (Athiwaratkun et al., 2023)	Multilingual programming benchmarks.
	MathQA-X (Athiwaratkun et al., 2023)	Multilingual MathQA for code reasoning.
	DS-1000 (Lai et al., 2023)	Data science programming benchmark.
	CodeXGLUE (Lu et al., 2021)	Code understanding and generation benchmark.
	CRUXEval (Gu et al., 2024)	Execution reasoning evaluation.
	LiveCodeBench (Jain et al., 2024)	Contamination-free code evaluation.
Dialogue & Multi-turn	**MT-Bench** (Zheng et al., 2023)	Multi-turn dialogue evaluation.
	MINT (Wang et al., 2024f)	Multi-turn instruction following.
	MuTual (Cui et al., 2020)	Multi-turn dialogue reasoning.
	DREAM (Sun et al., 2019)	Dialogue-based reading comprehension.
	QReCC (Anantha et al., 2021)	Question rewriting for conversational context.
	CICERO (Ghosal et al., 2022)	Conversational inference and causal reasoning.
	MUSR (Sprague et al., 2024)	Multi-turn soft reasoning.
	ClemBench (Chalamalasetti et al., 2023)	Game-theoretic rule adherence.