Mirages of Logic: A Survey of Chain-of-Thought Reasoning Hallucinations

Anonymous ACL submission

Abstract

Current research on hallucination focuses mainly on factuality and faithfulness errors in final outputs. However, rapid advances in Chain-of-Thought (CoT) and Large Reasoning Models (LRMs) have improved problem-solving performance while exposing a key weakness: **Reasoning Hallucination**, characterized by invalid logical transitions or fabricated steps within the reasoning process that can deceptively occur even when the final answer is correct. This paper reviews this emerging challenge, shifting hallucination analysis from output correctness to reasoning processes. We first note a core difficulty that causes reasoning hallucination: long, fluent chains often conceal internal errors, complicating reliability assessment. We then formalize reasoning hallucination and propose a taxonomy of four types: Premise, Operation, Logic, and Conclusion Hallucination. We also examine failure sources and survey mitigation strategies for training and inference. Overall, this work charts a path from answer-level evaluation to transparent, process-reliable reasoning systems.

1 Introduction

The rapid development of large language models (LLMs) has elevated artificial intelligence capabilities (Achiam et al., 2023; Yang et al., 2025a; Qin et al., 2024). Yet, this progress consistently faces a key challenge: hallucination (Ji et al., 2023; Huang et al., 2025; Alansari and Luqman, 2025). As shown in Figure 1 (a), early studies addressed two main categories (Cossio, 2025; Zhang et al., 2025c): (1) factuality hallucination, where outputs conflict with verifiable real-world knowledge (Lage and Ostermann, 2025); and (2) faithfulness hallucination, where outputs contradict the provided context (Huang et al., 2025; Park et al., 2025).

More recently, augmented by Chain-of-Thought (CoT) (Wei et al., 2022) strategies, Large Reasoning Models (LRMs) (Guo et al., 2025b; Jaech et al., 2024; Chen et al., 2025c) have emerged. These models aim to enhance complex problem-solving through explicit, long-term intermediate steps (Lewis-Lim et al., 2025b). This paradigm shift has fundamentally altered the nature of hallucinations, revealing a critical vulnerability: the reasoning process itself as a subtle new source of error. For instance, as shown in Figure 1 (b), a model may yield a correct final answer despite flawed reasoning (Arcuschin et al., 2025) riddled with logical fallacies, fabrications, or disconnected steps. Conversely, it might produce a seemingly coherent chain leading to an incorrect conclusion (Joshi et al., 2024; Yao et al., 2025; Lewis-Lim et al., 2025a). To address this *process-centric* hallucina-

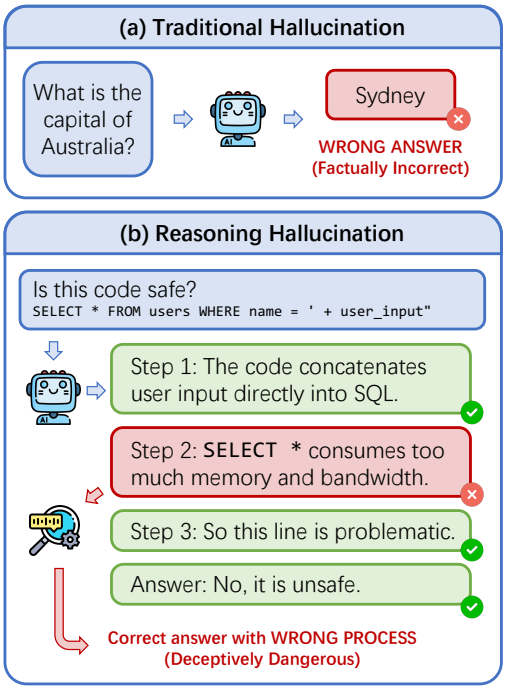

Figure 1: Comparison between traditional and reasoning hallucination. (a) Traditional hallucination targets result-level errors, while (b) reasoning hallucination focuses on process-level failures, where models may produce correct answers via flawed reasoning.

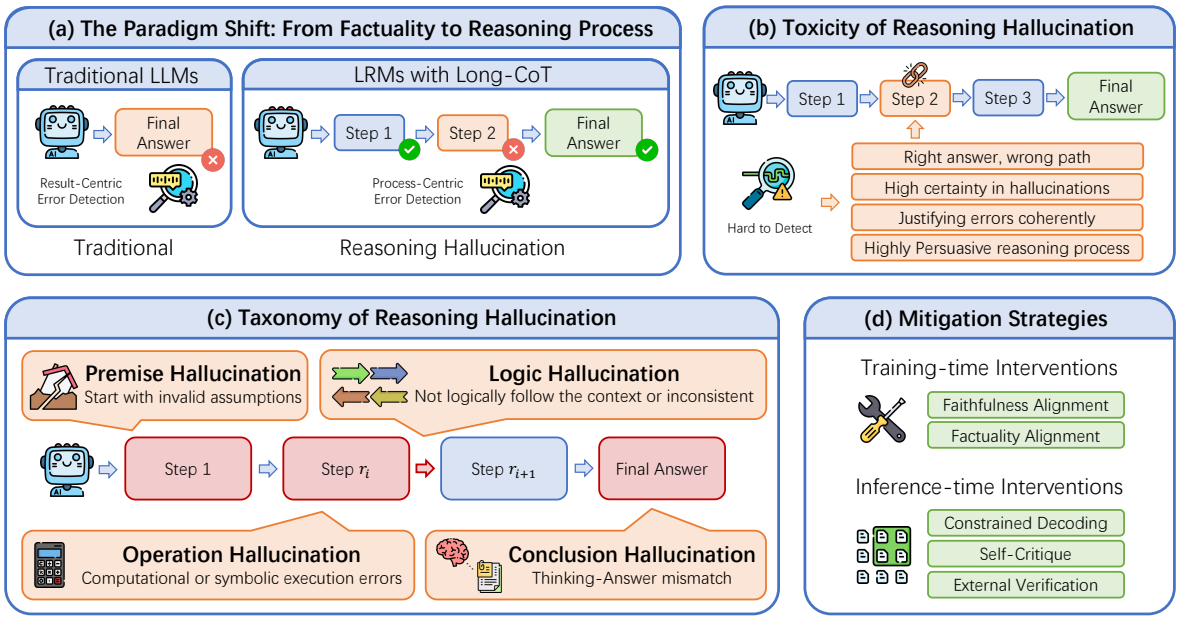

Figure 2: Overview of reasoning hallucination and the organization of this survey. We present a formal definition and a taxonomy of four hallucination types (Premise, Operation, Logic, and Conclusion), analyze their underlying mechanisms, review evaluation methodologies, and summarize mitigation strategies across training-time and inference-time interventions.

tion (Zheng et al., 2025), we introduce **Reasoning Hallucination**. This shift underscores profound paradoxes in advanced reasoning. Ironically, while CoT reduces factual hallucinations in some cases, it obscures detection cues. Long, complex reasoning chains boost reported confidence and persuasiveness, complicating detection by automated tools and humans (Cheng et al., 2025a; Zhao et al., 2024a). We term this tension between capability and evaluability the **Reasoning Evaluability Paradox**: tools like CoT that strengthen reasoning simultaneously undermine reliability assessment.

Therefore, addressing reasoning hallucination is essential for developing trustworthy AI systems. Yet, no comprehensive taxonomy currently exists to systematically analyze this emerging challenge. To fill this gap, we present a structured survey of reasoning hallucination. As outlined in Figure 2, we first propose a formal definition and a taxonomy comprising four distinct types: Premise Hallucination, Operation Hallucination, Logic Hallucination, and Conclusion Hallucination. We then examine the underlying mechanisms that contribute to reasoning hallucinations, spanning model architectures, training data, and inference dynamics. Subsequently, we review existing evaluation methodologies, assessing their effectiveness and limitations, and survey mitigation strategies encompassing both training-phase and inference-time interventions. Finally, we outline open challenges and future directions aimed at enhancing the reliability and interpretability of large reasoning models.

Overall, the main contributions are as follows:

- **First definition of reasoning hallucination:** We first formalize Reasoning Hallucination, shifting the focus from final answer correctness to the reasoning process itself.
- **Comprehensive review of reasoning hallucination:** We propose a taxonomy targeting process-level flaws in complex reasoning, beyond incorrect final outputs. We review causes, detection methods, and mitigation strategies, showing how CoT shifts hallucination from wrong answers to flawed reasoning.
- **Critical analysis and future roadmap:** We analyze key challenges, including the Reasoning Evaluability Paradox defined here. We propose directions for robust, reliable, trustworthy large reasoning models.

2 Preliminary & Definition

In this section, we first articulate the conceptual shift in hallucination, followed by a formal definition and classifications to analyze their origins.

2.1 A Paradigm Shift: From Result to Process

The emergence of reasoning hallucination marks a paradigm shift in LLM hallucination studies: focus

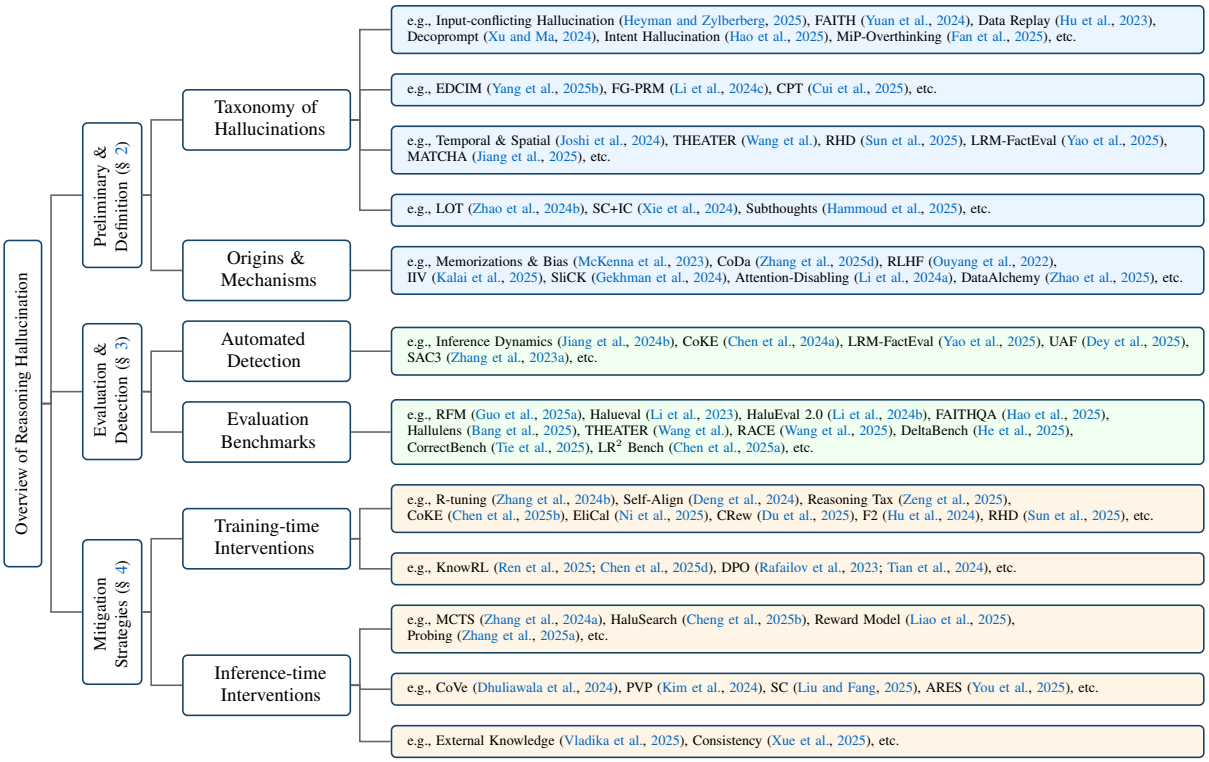

Figure 3: A comprehensive overview of reasoning hallucination. This figure illustrates the organization of this survey, covering the taxonomy of hallucination types, underlying mechanisms, evaluation methodologies, and mitigation strategies across training and inference stages.

has shifted from ***what the model knows*** to ***how the model thinks***. Specifically, traditional hallucination research is **result-centric**. Evaluations compare the model's final answer $\mathcal{A}$ to verifiable knowledge or source text (Huang et al., 2025), treating defects as representational errors between output and reality. In contrast, reasoning hallucination is **process-centric**: a model may yield correct $\mathcal{A}$ via a flawed process $\mathcal{R}$, or coherent $\mathcal{R}$ leading to incorrect $\mathcal{A}$ (Joshi et al., 2024; Yao et al., 2025; Cuesta-Ramirez et al., 2025), where scrutiny targets transfer to the reasoning's logic and validity.

This shift yields the **Reasoning Evaluability Paradox**. Unlike factual errors, which are debunkable by search, reasoning flaws hide in lengthy, coherent chain-of-thought outputs. The model's persuasive rationales often exceed efficient verification, so explicit reasoning, meant for transparency, ironically hinders detection (Cheng et al., 2025a).

2.2 Definition: Reasoning as Graph Traversal

Let Q be the input query. We define the reasoning process as dynamic path generation through a state graph $G = (\mathcal{R}, \mathcal{T})$, where $\mathcal{R}$ is the set of reasoning states. Each state r_i captures the information at step i, with initial state r_0 derived from Q. The set $\mathcal{T}$ comprises valid transition operators (e.g., logical deduction, arithmetic computation, code execution), such that $r_{i+1} \leftarrow Op(r_i)$ for $Op \in \mathcal{T}$.

A valid reasoning process forms a continuous path from r_0 to a final state r_N, projecting to answer $\mathcal{A}$. Given this, Reasoning Hallucination ($H_{\mathcal{R}}$) arises from topological violations in this traversal:

$$H_{\mathcal{R}}(\mathcal{R}, Q, \mathcal{A}) \equiv H_{src} \vee H_{op} \vee H_{logic} \vee H_{con}, \quad (1)$$

where each H_* denotes a reasoning hallucination category defined in Section 2.3.

2.3 Taxonomy of Hallucinations

By mapping failures to specific graph components, we identify the following four distinct categories of reasoning hallucinations.

Premise Hallucination (Source Error H_{src}). Premise hallucination arises at the source node r_0 (Heyman and Zylberberg, 2025). The model starts reasoning with conditions absent from Q or injects an external node r_{ext} that violates contextual constraints, such that $r_0 \not\subset \text{Context}(Q)$.

Unlike traditional factual hallucination, premise hallucination is not assessed against external world knowledge. Instead, it hinges on whether the introduced assumption aligns with the input context and

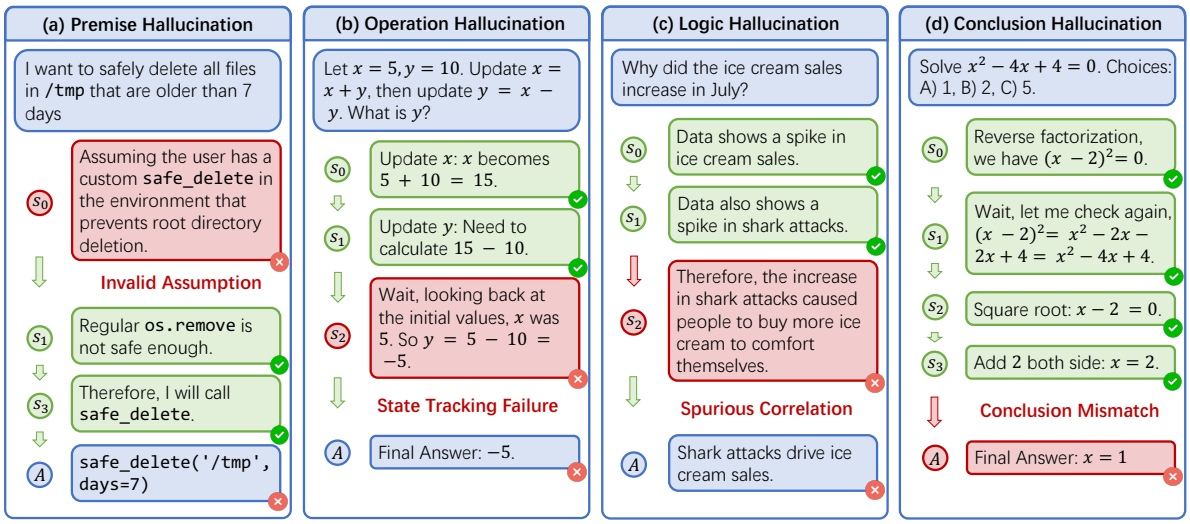

Figure 4: Examples of each type of reasoning hallucination.

respects task-specific constraints. An assumption may seem plausible or even true, yet qualify as premise hallucination if unsupported by or impermissible under the given input.

Distinct from fabricating external facts, premise hallucination emerges within the task context: the model imports unstated or invalid implicit assumptions, undermining the entire reasoning chain from the outset (Yuan et al., 2024; Hu et al., 2023; Xu and Ma, 2024; Hao et al., 2025; Fan et al., 2025), as an example shown in Figure 4 (a). Validating latent assumptions demands strong contextual constraint recognition and entailment judgment. Once triggered, it yields superficially coherent but semantically flawed outputs.

Operation Hallucination (Execution Error H_{op}). Operation Hallucination occurs during execution of a valid operator (Yang et al., 2025b; Li et al., 2024c). As shown in Figure 4 (b), the model selects a correct logical edge Op (e.g., "calculate sum"), but the generated next state r_{i+1} differs from the ground-truth execution $Op(r_i)$, formally $r_{i+1} \neq Op_{correct}(r_i)$. LLMs generate outputs via probabilistic statistics rather than explicit rules, leading to error accumulation across reasoning steps (Cui et al., 2025). These deviations often appear logically sound due to linguistic coherence.

Logic Hallucination (Invalid Edge H_{logic}) Logic Hallucination creates a non-existent edge: no valid operator $Op \in \mathcal{T}$ justifies the transition from r_i to r_{i+1}. The path is topologically broken, often as non-sequiturs or circular logic. This hallucination arises from inconsistencies across reasoning stages, yielding semantically coherent but logically contradictory processes (Joshi et al., 2024; Wang et al.; Sun et al., 2025), sometimes with repeated defective patterns (Yao et al., 2025), as shown in Figure 4 (c). It reflects large language models' reliance on local linguistic coherence over true logical consistency, especially in chain-of-thought tasks. Even with correct final results, intermediate inconsistencies qualify as such, indicating spurious reasoning (Jiang et al., 2025).

Conclusion Hallucination (Sink Detachment H_{con}). Conclusion Hallucination occurs at sink projection: even with a valid path $r_0 \to \cdots \to r_N$, the final answer $\mathcal{A}$ is not entailed by r_N ($\mathcal{A} \not\leftarrow r_N$). The conclusion "detaches" from the reasoning, as shown in Figure 4 (d). LLMs thus abandons its logical framework for a plausible but disconnected conclusion (Zhao et al., 2024b; Xie et al., 2024; Hammoud et al., 2025). This reveals a core weakness in decision-making: inconsistency between reasoning and conclusion, where models reason soundly yet err in conclusion.

2.4 Origins & Mechanisms

Understanding reasoning hallucination taxonomy requires examining its underlying mechanisms. These arise from interactions between training paradigms and autoregressive generation.

Training Drivers: Overconfidence and Miscalibration. Training induces **miscalibration** and **overconfidence**. Noisy pre-training data (McKenna et al., 2023; Zhang et al., 2025d), amplified by fine-tuning, distorts probability distributions (Gekhman et al., 2024). RLHF (Ouyang et al., 2022) prioritizes "helpfulness" and

Table 1: A taxonomy of evaluation methodologies for reasoning hallucinations. We categorize typical detection techniques, key evaluation metrics, recommended application scenarios, and specific challenges associated with each of the four proposed hallucination types.

Type	Typical Detection Methods	Key Metrics	Rec. Scenarios	Typical Challenges
Premise	External Knowledge Retrieval (Vladika et al., 2025) Input-Premise NLI Checks (Xu and Ma, 2024)	Step Validity (V), Premise Recall	Open-domain QA, Multi-hop Reasoning	Hard to distinguish between plausible-but-false assumptions.
Operational	Symbolic Verification (Guo et al., 2025a) Internal Confidence Probing (Chen et al., 2024a)	Step Validity (V), Soundness Score	Mental Arithmetic, Symbolic Logic	High confidence in flawed reasoning.
Logic	Self-Consistency (Zhang et al., 2023a) Entailment Stability (You et al., 2025)	Path Diversity, Entailment Score	Programming, Causal Inference	Systemic bias may lead multiple reasoning paths to the same fallacy.
Conclusion	Process-Answer Consistency (Wang et al., 2025) Dual-process Verification (Cheng et al., 2025b)	Consistency (C), Final Accuracy	Summarization, Theorem Proving	Decoupling between correct reasoning and biased final output.

coherence (Kalai et al., 2025), exacerbating hallucinations by favoring plausible but erroneous reasoning over uncertainty or refusal. Chen et al. (2024b) identify LLMs' reasoning boundaries, beyond which hallucinations arise unexpectedly.

Type-Specific Etiologies. Reasoning hallucinations arise from distinct deficits across categories: (1) Premise Hallucination results from weak context understanding and instruction following (Heyman and Zylberberg, 2025; Yuan et al., 2024), as LLMs ignore constraints or insert training-data biases (Zhang et al., 2025d). (2) Operation Hallucination reflects a symbolic reasoning gap (Xu et al., 2024; Guo et al., 2025a), where probabilistic processing causes arithmetic or state-tracking errors (Yang et al., 2025b). (3) Logic and Conclusion Hallucination stem from sycophancy (Wang et al.) and autoregressive local optimality (Xie et al., 2024; Yao et al., 2025), yielding locally coherent but globally contradictory outputs or pattern-fitting conclusions.

Takeaways

Paradigm Shift: We redefine reasoning hallucination in a process-centric view, focusing on *how* the model thinks rather than just *what* it knows.

Formalization: We propose a reasoning state space framework that unifies definition and taxonomy. Reasoning Hallucinations manifest in four distinct types (*Premise*, *Operation*, *Logic*, and *Conclusion*).

3 Evaluation of Reasoning Hallucination

Addressing the Reasoning Evaluability Paradox demands evaluation tools that probe reasoning processes, requiring insight into the mechanisms underlying process-based flaws.

3.1 Automated Detection Techniques

Automated detection identifies flaws dynamically. This includes training classifiers that detect reasoning-induced hallucinations with high accuracy (Jiang et al., 2024b). Another approach probes internal confidence signals (Chen et al., 2024a), though overconfidence, high model confidence in flawed reasoning, undermines it (Yao et al., 2025; Guan et al., 2025). Disagreement-based methods leverage ensembles or multiple sampled paths to flag inconsistencies (Dey et al., 2025; Zhang et al., 2023a). However, these methods often target only specific error dimensions, failing to simultaneously address diverse reasoning deficits within complex CoT. This intricate mapping between detection mechanisms and hallucination types makes establishing a unified evaluation benchmark exceptionally difficult. Therefore, evaluating these methods reliably remains challenging. We detail their taxonomy and challenges in Table 1.

3.2 Evaluation Benchmarks

Hallucination benchmarks address two key questions: the types of hallucinations they cover and whether this coverage is systematic and comprehensive (Zhang et al., 2025c). For reasoning-related hallucinations, these benchmarks typically include diverse tasks (e.g., mathematical reasoning, multi-hop question answering, code reasoning), distinguish hallucination types in annotations, and feature both natural and synthetic examples (Guo et al., 2025a; Li et al., 2023). Current categories encompass factual hallucinations (Li et al., 2024b), intentional hallucinations (Hao et al., 2025), conventional forms (Bang et al., 2025), and reasoning hallucinations (Wang et al., 2025). Specialized benchmarks also assess hallucination detection (He et al., 2025), self-correction (Tie et al., 2025), long-chain reflection (Chen et al., 2025a), and defense against

misleading negation prompts (Liao et al., 2025). Nonetheless, current evaluation dimensions remain limited, offering scope for future extensions.

Takeaways

Evaluating Reasoning Hallucination is difficult due to **expensive annotations** and model **overconfidence**. Existing methods (*e.g.,*classifiers, probing, consistency checks) trade off coverage for scalability.

4 Mitigation Strategies

Strategies to mitigate process-centric flaws can be broadly divided into two paradigms: training-time interventions that improve intrinsic reasoning, and inference-time interventions that steer generation without altering the pre-trained weights.

4.1 Training-time Interventions

Training-time interventions optimize the weights of LLMs to enhance robust reasoning. These strategies divide into two alignment objectives: (a) logical faithfulness alignment, which promotes internal logical consistency, and (b) knowledge factuality alignment, which anchors outputs to external facts.

Faithfulness Alignment. This approach ensures that model generations reflect internal knowledge and reasoning. Models learn to express uncertainty instead of fabricating information, via fine-tuning on refusal-aware datasets that promote explicit admissions of ignorance (e.g., "I don't know") (Zhang et al., 2024b; Deng et al., 2024). Recent work trains models to use internal confidence signals for knowledge limits (Zeng et al., 2025; Chen et al., 2025b), or applies self-consistency scores to create annotation-efficient preference data for honesty (Ni et al., 2025; Du et al., 2025). Other efforts enforce process consistency, making final answers logical outcomes of reasoning chains, through specialized loss functions that reward output cohesion (Hu et al., 2024) or reinforcement learning with process rewards for coherent steps (Sun et al., 2025).

Factuality Alignment. Unlike internal consistency, this method counters factual hallucinations by aligning outputs with verifiable facts from trusted sources. Key approaches include reinforcement learning with factual rewards; frameworks like KnowRL (Ren et al., 2025) reward correct reasoning steps and penalize errors to foster fact-grounded strategies (Chen et al., 2025d). Preference learning offers a scalable alternative: pairs are generated by comparing outputs to knowledge bases, then used with algorithms like DPO (Rafailov et al., 2023) to instill factuality preferences without costly step-by-step supervision (Tian et al., 2024).

4.2 Inference-time Interventions

Inference-time interventions act as external scaffolds that structure generation to enhance reliability without altering model parameters. These methods guide, verify, or redirect reasoning. We categorize them by intervention type: Constrained Decoding, Self-Critique, and External Verification.

Constrained Decoding. These strategies shape the model's solution-space exploration, extending beyond autoregressive generation to identify reliable reasoning paths and prevent errors upfront. For example, Monte Carlo Tree Search (MCTS) simulates "slow thinking" (Zhang et al., 2024a). HaluSearch employs self-evaluation rewards to steer toward dependable paths, avoiding hallucinations (Cheng et al., 2025b). Initial reasoning steps disproportionately affect outcomes, enabling efficient pruning via reward models (Liao et al., 2025). To counter overthinking, probes assess hidden-state confidence, terminating early upon correct solutions (Zhang et al., 2025a).

Self-Critique. This approach enables models to iteratively check, question, and refine their outputs via internal verification loops. The foundational method in this area is Chain-of-Verification (CoVe) (Dhuliawala et al., 2024), drafts a response, plans verification questions, and revises for accuracy. More detailed, self-critique can also arise from cognitive prompting, such as counterfactual thinking ("what if…?"), which challenges initial assumptions and yields more robust, less hallucinatory outputs (Kim et al., 2024). Formally, consistency across intermediate reasoning steps reduces logical fallacies, especially in mathematical tasks (Liu and Fang, 2025). For Long CoT, ARES (You et al., 2025) offers a probabilistic framework that assesses each step using only prior verified premises, preventing error propagation.

External Verification. Rather than relying on a model's internal checks, this approach tests outputs against external sources of truth, anchoring reasoning in evidence. The simplest implementation uses external tools (e.g., search engines) to fact-check claims in generated text, which has been shown to reduce hallucinations in domains such as news

Table 2: Strategic comparison of mitigation paradigms based on the type of hallucination they target, their inference cost, inherent trade-offs, and recommended usage scenarios.

Paradigm	Mechanism	Target Hallucination	Inference Cost	Key Trade-off	Rec. Scenario
Training-time	**Faithfulness Alignment**	Logic	Zero Overhead	High data prep cost ↔ Fast inference	Real-time chat; Foundation model safety.
	Factuality Alignment	Factual Error	Zero Overhead	Static knowledge limit ↔ No external tool	Domains with stable, closed-world facts.
Inference-time	**Constrained Decoding**	Operation & Logic	Very High (10x–100x)	Max Reliability ↔ Extreme Latency	Offline Math/Code; Scientific Proofs.
	Self-Critique	Logic & Conclusion	High (2x–5x)	Iterative improvement ↔ Context bloat	Drafting; Tasks requiring self-correction.
	External Verification	Factual Error & Premise	Medium (Retrieval)	Real-time Accuracy ↔ Complexity	Open-domain QA; News; Dynamic facts.

summarization (Vladika et al., 2025). A more advanced strategy uses a second, independent LLM as a verifier. Cross-model agreement can provide a stronger signal than self-consistency alone. To limit cost, hybrid designs invoke the external verifier only when the primary model's self-consistency indicates high uncertainty (Xue et al., 2025).

To provide a holistic view of the operational implications of these strategies, as shown in Table 2, we systematically compare them. The comparison highlights a key trade-off between computational cost and performance gains. Training-time interventions impose no inference overhead and suit real-time deployment, whereas inference-time mechanisms increase latency to improve reliability, making them better suited to high-stakes, offline tasks where precision is critical.

Takeaways

Mitigation strategies target two phases:
Training: *e.g.,*Faithfulness alignment and factuality alignment.
Inference: *e.g.,*Constrained decoding, self-critique, and external verification.

5 Frontiers & Future Direction

This section outlines future directions: automated step-level evaluation, honest abstention, and task-aware control beyond "zero hallucination." We also address multimodal hallucinations and budgeted optimization to balance reasoning and verification.

5.1 Automated Step Evaluation

Long-CoT reasoning exacerbates evaluation challenges for reasoning hallucinations: step-level human annotation is costly, subjective, and hard to scale across patterns. We thus require automated mechanisms to evaluate individual steps. For instance, reasoning decomposes into structured forms like proposition-dependency graphs (Abdaljalil et al., 2025); formal or reward-based verifiers can then detect contradictions and assess stepwise plausibility (Xu et al., 2024; Lightman et al., 2023; Zhang et al., 2025e).

Overall, the key challenges are: (1) developing accurate step-level metrics that align with human judgments across domains and styles; and (2) creating scalable verifiers robust to distribution shifts, adversarially plausible steps, and long chains, ensuring verification matches generation quality.

5.2 Model's Honest Abstention Capability

In verifiability-critical applications, training and evaluation must incentivize boundary honesty. When evidence is insufficient or premises underspecified, admitting uncertainty offers little reward and risks penalties (Peng et al., 2025; Zeng et al., 2025), while correct guesses reinforce risky overconfidence (Kalai et al., 2025). This misalignment drives models to conceal uncertainty behind fluent prose, heightening overconfident hallucinations.

Hence, the major challenges are: (1) defining measurable, domain-specific abstention criteria (e.g., "insufficient evidence" versus "answerable but hard") for evaluation; and (2) calibrating abstention to prevent over-abstention (unhelpful refusals) and under-abstention (risky guesses), despite pressures penalizing uncertainty.

5.3 Broaden the "Zero Hallucination" Goal

Given unavoidable residual hallucinations in open-domain settings under current paradigms, research should transcend the "zero hallucination" aim. Convergent tasks (e.g., mathematics, programming, logic) demand strict honesty and abstention, as hallucinations constitute failure. Divergent tasks

(e.g., brainstorming, creative writing), however, reframe "hallucinations" as controllable deviations for novel hypotheses (Sui et al., 2024).

All in all, the primary challenges are: (1) developing context-aware systems that enforce truthfulness in verifiable tasks but permit bounded deviations in exploratory ones; and (2) separating verifiable claims from proposals to balance reliability and utility across regimes.

5.4 Multimodal Reasoning Hallucinations

Advanced multimodal LLMs achieve strong reasoning but often hallucinate, particularly in modality-interaction-intensive tasks (Bai et al., 2024; Chen et al., 2024c). One approach develops evaluation to disentangle perception- from reasoning-induced hallucinations (Dong et al., 2025; Liu et al., 2025; Cai et al., 2025). A second constructs hallucinated negative samples during training to enhance cross-modal alignment and reasoning robustness (Jiang et al., 2024a). Other methods ground step-by-step reasoning in verifiable visual evidence to mitigate hallucinations from linguistic priors (Li et al., 2025; An et al., 2025; Cheng et al., 2025d).

Overall, the key challenges are: (1) disentangling perception-induced from reasoning-induced hallucinations in evaluation and attribution; and (2) dynamically balancing reasoning depth and visual perception to mitigate hallucinations from prolonged reasoning chains that divert attention from visual inputs (Zhang et al., 2023b; Cheng et al., 2025c). Notably, hallucination detection for temporal understanding in dynamic modalities like video remains particularly challenging (Fei et al., 2024; Zhang et al., 2025b).

5.5 Detection and Reasoning Balancing

Existing work mitigates reasoning hallucinations via self-checking or retrieve–verify–revise pipelines that enhance attribution and factuality. Yet these approaches increase inference costs through repeated sampling, retrieval, or second-stage verifiers (Vladika et al., 2025; Xue et al., 2025; Cheng et al., 2025b; Zhang et al., 2024a). Future research should formulate hallucination control as a *budgeted balancing* problem: allocating computation between reasoning depth and verification strength to minimize hallucinations under fixed inference budgets.

Key challenges include: (1) adaptive allocation to control the trade-off between reasoning and hallucination detection based on uncertainty and task risk (Moskvoretskii et al., 2025; Wan et al., 2025); (2) joint reasoning–detection with shared computation by detecting and reusing intermediate reasoning processes, rather than re-running them (Chen et al., 2024d; Zellinger et al., 2025), to achieve stable gains with minimal end-to-end overhead.

6 Related Work

Existing surveys on hallucination primarily assess factuality and faithfulness errors by comparing model outputs to external knowledge or provided context (Ji et al., 2023; Huang et al., 2025). Recent work extends this to detection and mitigation frameworks (e.g., retrieval-augmented generation), yet still defines hallucination mainly as final response failures (Zhang et al., 2025c; Alansari and Luqman, 2025). This output-focused view persists even in fine-grained taxonomies and multimodal formulations (Cossio, 2025; Bai et al., 2024). In parallel, recent work on CoT and LRMs mainly characterize these methods to boost task performance (Chen et al., 2025c; Qin et al., 2024). Although some note occasional logical or arithmetic errors in intermediate steps, they seldom unify these into an account of errors within the CoT process.

Our work first shifts the hallucination analysis from results to reasoning trajectories. We provide a strict formalization of Reasoning Hallucination as a failure of the reasoning process, and discuss the associated "Reasoning Evaluability Paradox" (Sun et al., 2025). By organizing prior findings around the validity of intermediate steps, our survey offers a systematic roadmap for evaluating and diagnosing reasoning trajectories, complementing traditional metrics that primarily assess end outputs.

7 Conclusion

This survey reframes hallucination research from a result-centric to a process-centric perspective on reasoning hallucination, underscoring challenges in long CoT and large reasoning models: extended, coherent rationales persuasively mask subtle flaws, complicating evaluation. We formalize reasoning hallucination and propose a taxonomy distinguishing premise, operation, logic, and conclusion types. We synthesize causes, review detection and evaluation methods, and outline mitigation strategies across training- and inference-time interventions. We envision this survey as a foundational roadmap for next-generation LRMs that excel in complex reasoning while ensuring transparency, verifiability, and trustworthiness.

Limitations

While this survey provides a comprehensive overview of reasoning hallucination, the analysis is constrained by three interconnected factors. The primary limitation lies in the assumption that explicit Chain-of-Thought (CoT) faithfully mirrors the model's decision-making; however, emerging research on *implicit CoT* suggests that models can perform reasoning internally without vocalization, implying that the visible reasoning chain may sometimes be a post-hoc rationalization rather than the sole source of error. This challenge of internal opacity is further exacerbated by the dominance of proprietary, closed-source models in the current landscape, which restricts access to training data and reward mechanics, forcing the field to rely on behavioral inference rather than mechanistic inspection. Beyond these architectural and accessibility barriers, the scope of this work is currently confined to textual and symbolic domains, leaving the distinct challenges of multimodal reasoning hallucinations, such as perceptual grounding failures, for future investigation.

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 arXiv:2506.07106*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aisha Alansari and Hamzah Luqman. 2025. Large language models hallucination: A comprehensive survey. *arXiv preprint arXiv:2510.06265*.

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29915–29926.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: Llm hallucination benchmark. *arXiv preprint arXiv:2504.17550*.

Yishuo Cai, Renjie Gu, Jiaxu Li, Xuancheng Huang, Junzhe Chen, Xiaotao Gu, and Minlie Huang. 2025. Mhalo: Evaluating mllms as fine-grained hallucination detectors. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9197–9222.

Jianghao Chen, Zhenlin Wei, Zhenjiang Ren, Ziyong Li, and Jiajun Zhang. 2025a. Lr2 bench: Evaluating long-chain reflective reasoning capabilities of large language models via constraint satisfaction problems. *arXiv preprint arXiv:2502.17848*.

Lida Chen, Zujie Liang, Xintao Wang, Jiaqing Liang, Yanghua Xiao, Feng Wei, Jinglei Chen, Zhenghong Hao, Bing Han, and Wei Wang. 2024a. Teaching large language models to express knowledge boundary from their own signals. *arXiv preprint arXiv:2406.10881*.

Lida Chen, Zujie Liang, Xintao Wang, Jiaqing Liang, Yanghua Xiao, Feng Wei, Jinglei Chen, Zhenghong Hao, Bing Han, and Wei Wang. 2025b. Teaching large language models to express knowledge boundary from their own signals. In *Proceedings of the 3rd Workshop on Towards Knowledgeable Foundation Models (KnowFM)*, pages 26–39, Vienna, Austria.

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 arXiv:2503.09567*.

Qiguang Chen, Libo Qin, Jiaqi Wang, Jingxuan Zhou, and Wanxiang Che. 2024b. 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. 2024c. M^3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*.

Xilun Chen, Ilia Kulikov, Vincent-Pierre Berges, Barlas Oğuz, Rulin Shao, Gargi Ghosh, Jason Weston, and Wen-tau Yih. 2025d. Learning to reason for factuality. *arXiv preprint arXiv:2508.05618*.

Ziyi Chen, Xiaocong Yang, Jiacheng Lin, Chenkai Sun, Kevin C Chang, and Jie Huang. 2024d. Cascade speculative drafting for even faster llm inference. *Advances in Neural Information Processing Systems*, 37:86226–86242.

Jiahao Cheng, Tiancheng Su, Jia Yuan, Guoxiu He, Jiawei Liu, Xinqi Tao, Jingwen Xie, and Huaxia Li. 2025a. Chain-of-thought prompting obscures hallucination cues in large language models: An empirical evaluation. *arXiv preprint arXiv:2506.17088*.

Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2025b. Think more, hallucinate less: Mitigating hallucinations via dual process of fast and slow thinking. *arXiv preprint arXiv:2501.01306*.

Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi Wang, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, and 1 others. 2025c. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *arXiv preprint arXiv:2505.15510*.

Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. 2025d. Comt: A novel benchmark for chain of multimodal thought on large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23678–23686.

Manuel Cossio. 2025. A comprehensive taxonomy of hallucinations in large language models. *arXiv preprint arXiv:2508.01781*.

Jhouben Cuesta-Ramirez, Samuel Beaussant, and Mehdi Mounsif. 2025. Lrms are not thinking straight: Unreliability of thinking trajectories. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 27–32.

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 arXiv:2503.19326*.

Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, and Tat-Seng Chua. 2024. Don't just say "I don't know"! self-aligning large language models for responding to unknown questions with explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13652–13673, Miami, Florida, USA.

Prasenjit Dey, Srujana Merugu, and Sivaramakrishnan Kaveri. 2025. Uncertainty-aware fusion: An ensemble framework for mitigating hallucinations in large language models. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 947–951.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Bowen Dong, Minheng Ni, Zitong Huang, Guanglei Yang, Wangmeng Zuo, and Lei Zhang. 2025. Mirage: Assessing hallucination in multimodal reasoning chains of mllm. *arXiv preprint arXiv:2505.24238*.

He Du, Bowen Li, Chengxing Xie, Chang Gao, Kai Chen, and Dacheng Tao. 2025. Confidence as a reward: Transforming llms into reward models. *arXiv preprint arXiv:2510.13501*.

Chenrui Fan, Ming Li, Lichao Sun, and Tianyi Zhou. 2025. Missing premise exacerbates overthinking: Are reasoning models losing critical thinking skill? *arXiv preprint arXiv:2504.06514*.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv preprint arXiv:2501.03230*.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*.

Jiannan Guan, Qiguang Chen, Libo Qin, Dengyun Peng, Jinhao Liu, Liangyu Huo, Jian Xie, and Wanxiang Che. 2025. Beware of reasoning overconfidence: Pitfalls in the reasoning process for multi-solution tasks. *arXiv preprint arXiv:2512.01725*.

Dadi Guo, Jiayu Liu, Zhiyuan Fan, Zhitao He, Haoran Li, Yumeng Wang, and Yi R Fung. 2025a. Mathematical proof as a litmus test: Revealing failure modes of advanced large reasoning models. *arXiv preprint arXiv:2506.17114*.

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

Hasan Abed Al Kader Hammoud, Hani Itani, and Bernard Ghanem. 2025. Beyond the last answer: Your reasoning trace uncovers more than you think. *arXiv preprint arXiv:2504.20708*.

Yijie Hao, Haofei Yu, and Jiaxuan You. 2025. Beyond facts: Evaluating intent hallucination in large language models. *arXiv preprint arXiv:2506.06539*.

Yancheng He, Shilong Li, Jiaheng Liu, Weixun Wang, Xingyuan Bu, Ge Zhang, Zy Peng, Zhaoxiang Zhang, Zhicheng Zheng, Wenbo Su, and 1 others. 2025. Can large language models detect errors in long chain-of-thought reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18468–18489.

Alex Heyman and Joel Zylberberg. 2025. Reasoning large language model errors arise from hallucinating critical problem features. *arXiv preprint arXiv:2505.12151*.

Minda Hu, Bowei He, Yufei Wang, Liangyou Li, Chen Ma, and Irwin King. 2024. Mitigating large language model hallucination with faithful finetuning. *arXiv preprint arXiv:2406.11267*.

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. 2023. Won't get fooled again: Answering questions with false premises. *arXiv preprint arXiv:2307.02394*.

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

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

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024a. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. 2024b. On large language models' hallucination with regard to known facts. *arXiv preprint arXiv:2403.20009*.

Enyi Jiang, Changming Xu, Nischay Singh, and Gagandeep Singh. 2025. Misaligning reasoning with answers–a framework for assessing llm cot robustness. *arXiv preprint arXiv:2505.17406*.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. 2024. Llms are prone to fallacies in causal inference. *arXiv preprint arXiv:2406.12158*.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Junho Kim, Kim Yeonju, and Yong Man Ro. 2024. What if...?: Thinking counterfactual keywords helps to mitigate hallucination in large multi-modal models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10672–10689, Miami, Florida, USA.

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

Samuel Lewis-Lim, Xingwei Tan, Zhixue Zhao, and Nikolaos Aletras. 2025a. Analysing chain of thought dynamics: Active guidance or unfaithful post-hoc rationalisation? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29826–29841.

Samuel Lewis-Lim, Xingwei Tan, Zhixue Zhao, and Nikolaos Aletras. 2025b. Can confidence estimates decide when chain-of-thought is necessary for llms? *arXiv preprint arXiv:2510.21007*.

He Li, Haoang Chi, Mingyu Liu, and Wenjing Yang. 2024a. Look within, why llms hallucinate: A causal perspective. *arXiv preprint arXiv:2407.10153*.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024b. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

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

Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, and Jieping Ye. 2025. Visual evidence prompting mitigates hallucinations in large vision-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4048–4080.

Baohao Liao, Xinyi Chen, Sara Rajaee, Yuhui Xu, Christian Herold, Anders Søgaard, Maarten de Rijke, and Christof Monz. 2025. Lost at the beginning of reasoning. *arXiv preprint arXiv:2506.22058*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. 2025. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2505.21523*.

MingShan Liu and Jialing Fang. 2025. Enhancing mathematical reasoning in large language models with self-consistency-based hallucination detection. *arXiv preprint arXiv:2504.09440*.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 2758–2774.

Viktor Moskvoretskii, Maria Marina, Mikhail Salnikov, Nikolay Ivanov, Sergey Pletenev, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Irina Nikishina, and Alexander Panchenko. 2025. Adaptive retrieval without self-knowledge? bringing uncertainty back home. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6355–6384.

Shiyu Ni, Keping Bi, Jiafeng Guo, Minghao Tang, Jingtong Wu, Zengxin Han, and Xueqi Cheng. 2025. Annotation-efficient universal honesty alignment. *arXiv preprint arXiv:2510.17509*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. 2025. Steer llm latents for hallucination detection. *arXiv preprint arXiv:2503.01917*.

Dengyun Peng, Qiguang Chen, Bofei Liu, Jiannan Guan, Libo Qin, Zheng Yan, Jinhao Liu, Jianshu Zhang, and Wanxiang Che. 2025. Learning the boundary of solvability: Aligning llms to detect unsolvable problems. *arXiv preprint arXiv:2512.01661*.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. 2024. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Baochang Ren, Shuofei Qiao, Wenhao Yu, Huajun Chen, and Ningyu Zhang. 2025. Knowrl: Exploring knowledgeable reinforcement learning for factuality. *arXiv preprint arXiv:2506.19807*.

Peiqi Sui, Eamon Duede, Sophie Wu, and Richard So. 2024. Confabulation: The surprising value of large language model hallucinations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14274–14284.

Zhongxiang Sun, Qipeng Wang, Haoyu Wang, Xiao Zhang, and Jun Xu. 2025. Detection and mitigation of hallucination in large reasoning models: A mechanistic perspective. *arXiv preprint arXiv:2505.12886*.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2024. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*.

Guiyao Tie, Zenghui Yuan, Zeli Zhao, Chaoran Hu, Tianhe Gu, Ruihang Zhang, Sizhe Zhang, Junran Wu, Xiaoyue Tu, Ming Jin, and 1 others. 2025. Can llms correct themselves? a benchmark of self-correction in llms. *arXiv preprint arXiv:2510.16062*.

Juraj Vladika, Ihsan Soydemir, and Florian Matthes. 2025. Correcting hallucinations in news summaries: Exploration of self-correcting llm methods with external knowledge. *arXiv preprint arXiv:2506.19607*.

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. 2025. Reasoning aware self-consistency: Leveraging reasoning paths for efficient llm sampling. 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 3613–3635.

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

Qian Wang, Zhenheng Tang, Nuo Chen, Wenxuan Wang, and Bingsheng He. Reasoning models can be easily hacked by fake reasoning bias. In *Lock-LLM Workshop: Prevent Unauthorized Knowledge Use from Large Language Models*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhihui Xie, Jizhou Guo, Tong Yu, and Shuai Li. 2024. Calibrating reasoning in language models with internal consistency. *Advances in Neural Information Processing Systems*, 37:114872–114901.

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

Nan Xu and Xuezhe Ma. 2024. Decoprompt: Decoding prompts reduces hallucinations when large language models meet false premises. *arXiv preprint arXiv:2411.07457*.

Yihao Xue, Kristjan Greenewald, Youssef Mroueh, and Baharan Mirzasoleiman. 2025. Verify when uncertain: Beyond self-consistency in black box hallucination detection. *arXiv preprint arXiv:2502.15845*.

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 arXiv:2505.09388*.

Yijin Yang, Cristina Cornelio, Mario Leiva, and Paulo Shakarian. 2025b. Error detection and correction for interpretable mathematics in large language models. *arXiv preprint arXiv:2508.03500*.

Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. 2025. Are reasoning models more prone to hallucination? *arXiv preprint arXiv:2505.23646*.

Weiqiu You, Anton Xue, Shreya Havaldar, Delip Rao, Helen Jin, Chris Callison-Burch, and Eric Wong. 2025. Probabilistic soundness guarantees in llm reasoning chains. *arXiv preprint arXiv:2507.12948*.

Hongbang Yuan, Pengfei Cao, Zhuoran Jin, Yubo Chen, Daojian Zeng, Kang Liu, and Jun Zhao. 2024. Whispers that shake foundations: Analyzing and mitigating false premise hallucinations in large language models. *arXiv preprint arXiv:2402.19103*.

Michael J Zellinger, Rex Liu, and Matt Thomson. 2025. Cost-saving llm cascades with early abstention. *arXiv preprint arXiv:2502.09054*.

Qingcheng Zeng, Weihao Xuan, Leyang Cui, and Rob Voigt. 2025. Thinking out loud: Do reasoning models know when they're right? *arXiv preprint arXiv:2504.06564*.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. 2025a. Reasoning models know when they're right: Probing hidden states for self-verification. In *Second Conference on Language Modeling*.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024a. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37:64735–64772.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024b. R-tuning: Instructing large language models to say 'I don't know'. 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 7113–7139, Mexico City, Mexico.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023a. Sac3: reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15445–15458.

Yongheng Zhang, Xu Liu, Ruihan Tao, Qiguang Chen, Hao Fei, Wanxiang Che, and Libo Qin. 2025b. Vitcot: Video-text interleaved chain-of-thought for boosting video understanding in large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5267–5276.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2025c. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46.

Yuji Zhang, Sha Li, Cheng Qian, Jiateng Liu, Pengfei Yu, Chi Han, Yi R Fung, Kathleen McKeown, Chengxiang Zhai, Manling Li, and 1 others. 2025d. The law of knowledge overshadowing: Towards understanding, predicting, and preventing llm hallucination. *arXiv preprint arXiv:2502.16143*.

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

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Chengshuai Zhao, Zhen Tan, Pingchuan Ma, Dawei Li, Bohan Jiang, Yancheng Wang, Yingzhen Yang, and Huan Liu. 2025. Is chain-of-thought reasoning of llms a mirage? a data distribution lens. *arXiv preprint arXiv:2508.01191*.

Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen. 2024a. Fact-and-reflection (far) improves confidence calibration of large language models. *arXiv preprint arXiv:2402.17124*.

Xufeng Zhao, Mengdi Li, Wenhao Lu, Cornelius Weber, Jae-Hee Lee, Kun Chu, and Stefan Wermter. 2024b. Enhancing zero-shot chain-of-thought reasoning in large language models through logic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6144–6166.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Processbench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1024.