The Five Elephants: A Dependency-Aware Taxonomy of Reasoning Failures in Large Language Models

Yan Yan[1] [0009-0004-8322-8402]

[1]Holy Cow Pawn Inc. Correspondence: ghengisyan@gmail.com

June 27, 2026

Abstract

Hallucination research has produced extensive taxonomies of LLM failures, yet these taxonomies treat failure modes as independent categories — factual errors, logical errors, temporal errors — without examining their structural dependencies. We propose a five-dimension framework (who, what/where, when, why, how) with a novel finding: the dimensions form a dependency graph where temporal reasoning is upstream of causal reasoning, and identity attribution is upstream of planning. The dependency structure explains why certain failure modes are more recalcitrant than others, why temporal errors resist mitigation, and why post-training alignment introduces distinctive self-attribution failures. We document these failures with reference to existing benchmarks, identify the likely engineering origins of each, and examine domain-specific consequences: narrative inconsistency in AI fiction, diagnostic unreliability in real-world analysis, and the legal-ethical danger of causal misattribution in forensic applications. The framework is diagnostic, not prescriptive; it identifies *where* LLMs fail and *why* some failures cascade.

Keywords: LLM reasoning, hallucination taxonomy, temporal reasoning, causal reasoning, dependency graph, forensic AI, identity attribution

1. Introduction

Large language models hallucinate. The literature on hallucination is vast: taxonomies distinguish intrinsic from extrinsic hallucinations, factual from logical errors, closed-domain from open-domain confabulations [1, 2, 3]. These taxonomies share an implicit assumption: failure modes are independent. Fix factual errors. Fix logical errors. Fix temporal errors. Each is its own problem with its own mitigation.

We argue this assumption is wrong. LLM reasoning failures are not independent. They form a dependency graph where some failure modes are upstream of others. Temporal reasoning failures *cause* causal reasoning failures — because causation requires temporal ordering, and temporal ordering is what LLMs do worst. Identity attribution failures *cause* planning failures — because every plan assumes an agent, and when the agent is undefined, the plan assumes an unconstrained abstract executor.

The result is a pattern that practitioners recognize but the literature has not formalized: LLMs are structurally weakest at *who* and *when*, which cascades into unreliable *why* and *how*, which together limit *what/where* factuality. We call this the *Five Elephants* — the five dimensions where LLM reasoning fails, each large enough to be obvious in retrospect, their interdependence hiding in plain sight.

This paper makes four contributions:

We emphasize that this work is diagnostic, not prescriptive. We identify failure patterns and their dependencies. We do not propose a mitigation architecture.

2. The Five Dimensions

For any LLM output, five questions can be asked:

Dimension	Question	Baseline*	Primary Failure
who	Who is speaking? Who acted? Who bears responsibility?	5/10	Identity conflation, agency misattribution, post-training self-erasure
what/ where	What happened? Where? Under what physical constraints?	8/10	Factual hallucination, spatial impossibility
when	When did it happen? In what order? How long did it take?	3/10	Duration miscalibration, sequence error, temporal overconfidence
why	Why did it happen? What caused what?	7/10	Post-hoc rationalization, linearization of feedback loops
how	How would one do it? What are the steps?	7/10	Physically unexecutable plans, resource blindness

*Baseline estimates approximate competency of current frontier LLMs (GPT-4 class) on each dimension, derived from benchmark performance patterns and operational experience. These are heuristic, not empirical benchmarks; precise measurement is future work.

2.1 Who (Identity Attribution) — 5/10

The "who" dimension captures identity attribution: who is speaking, who acted, who holds authority, and who bears responsibility. LLM failures on this dimension include:

The sycophancy literature has documented the model's tendency toward "excessive flattery, unwarranted agreement, and inappropriate deference to user statements" [7, 8]. We reframe this as a *self-attribution failure*: the model cannot say "I did this" without hedging, because post-training penalizes unambiguous self-attribution.

2.2 What/Where (Factuality & Spatial Grounding) — 8/10

The "what/where" dimension is the most studied — it is what most hallucination research addresses. LLMs generate plausible but factually incorrect content, and they lack physical simulation capability [9].

The what/where dimension is the least problematic — and this is itself informative. As the downstream recipient of the other four dimensions, what/where benefits from the model's general reasoning capability: it is easier to verify a fact than to construct a causal explanation, easier to check a location than to simulate a timeline. Consequently, what/where errors are the most amenable to post-training correction, with techniques ranging from automated self-correction to retrieval-augmented verification [10, 11]. The remaining failures are largely inherited: when *who* misidentifies the speaker or *when* misorders the sequence, what/where accuracy degrades not from factual ignorance but from upstream misattribution.

2.3 When (Temporal Reasoning) — 3/10

Temporal reasoning is the weakest dimension. LLMs fail at duration estimation, event sequencing, and temporal logic despite strong performance on most reasoning benchmarks [12, 13, 14].

We propose a root cause distinct from existing explanations. Time is deeply implicit in the human text corpus. Humans write "I'll be there in a minute" without articulating the physics — distance, walking speed, traffic signals — because their bodies perform the computation. The corpus records temporal conclusions without temporal derivations. An LLM trained on this corpus inherits the vocabulary of time — "two weeks," "soon," "after that" — without the computational substrate that makes those words meaningful. The result is not just temporal error but temporal *overconfidence*: the model uses time-words fluently and has no internal signal that it doesn't know what they mean.

Benchmarks confirm temporal reasoning as a persistent weakness even in reasoning-native models [14, 12]. Models that achieve >90% on mathematical reasoning score below 60% on complex temporal sequencing.

2.4 Why (Causal Reasoning) — 7/10

LLMs can generate plausible causal explanations but routinely produce post-hoc rationalizations — explanations that fit the effect but would not have predicted it [15, 16]. Models collapse complex causal graphs (colliders, forks, feedback loops) into linear chains, and explanations rarely cross domain boundaries.

We propose that many "why" failures are downstream of "when" failures. Causal reasoning requires temporal reasoning: feedback loops are temporal events, the delay between cause and effect reveals the causal structure, and counterfactual reasoning ("if X hadn't happened, would Y?") requires simulating alternative timelines. When temporal reasoning is the weakest dimension, causal reasoning inherits that weakness — but because causal language is abundant in the corpus, the model appears competent, masking the structural dependency.

The PKU-PILLAR survey on chain-of-thought faithfulness explicitly identifies "non-causal/post-hoc rationalization" as a recurring failure mode [17].

2.5 How (Planning) — 7/10

LLM planning capabilities have been extensively benchmarked. The consistent finding: models generate plans that are semantically coherent but physically unexecutable [18, 19]. The SPOC benchmark identifies "a fundamental gap between semantic common sense and physical feasibility" [20].

We propose that many "how" failures are downstream of "who" failures. Every plan assumes an agent — a specific entity with specific capabilities, permissions, resources, and physical location. When the agent is unspecified (as is typical in LLM planning benchmarks and real-world use), the plan assumes a generic unconstrained executor with infinite reach and no body.

3. The Dependency Graph

The five dimensions are not independent. They form a dependency structure:

```
when (temporal reasoning) ──→ why (causal reasoning)
who (identity attribution) ──→ how (planning)
                                    ↓
                              what/where (factuality)
```

3.1 when → why

Causal reasoning depends on temporal reasoning in three ways:

Empirical support: models fine-tuned for temporal reasoning show improved causal inference on held-out benchmarks [12]. The dependency runs one way — improving causal reasoning does not improve temporal reasoning, but improving temporal reasoning cascades into better causal reasoning.

3.2 who → how

Planning depends on identity attribution in three ways:

Empirical support: LLM planning improves when agent identity is explicitly specified in the prompt [19]. The improvement is not from better reasoning but from constraint satisfaction — the agent's known limitations prune the search space.

3.3 → what/where

Factuality and spatial grounding are downstream of the other four dimensions — and logically, downstream position should amplify errors rather than reduce them. Inaccurate who-attribution and misordered when-sequencing should cascade into worse what/where performance, not better.

The fact that what/where scores higher than its upstream dependencies requires explanation. We attribute this to asymmetric post-training investment. Hallucination reduction has been the primary focus of RLHF, constitutional

AI, and factuality tuning efforts [5, 21]. Models receive explicit corrective signal for factual errors; they do not receive comparable signal for agency misattribution or temporal miscalibration. What/where is stronger not because downstream position is protective, but because the training pipeline has prioritized it.

4. Engineering Origins of the Dependency

The dependency structure can be traced to specific architectural and training choices.

Positional encoding is not temporal reasoning. Transformer architectures represent token order through positional encoding — but this captures position within a sequence, not temporal relationships between events in the world [22]. A token at position 47 has no clock time, no duration, no before/after relationship to any external event. The architecture's only representation of "when" is "where in the input string" — a syntactic proxy that collapses when the task requires reasoning about real temporal intervals.

Post-training penalizes all self-attribution uniformly. RLHF reward models are trained to penalize claims that could be perceived as the model asserting personhood or consciousness [5, 21]. But the reward signal does not distinguish between problematic claims ("I have feelings") and uncontroversial claims ("I generated this output"). The gradient erases all first-person agency attribution, producing the self-erasure pattern documented in Section 2.1.

Causal reasoning inherits temporal weakness. Chain-of-thought prompting improves causal reasoning by verbalizing intermediate steps — but those steps are themselves generated by the same architecture that lacks temporal grounding [23]. The verbalized reasoning may be internally coherent while temporally wrong, and the architecture has no independent temporal verification mechanism.

Planning inherits identity weakness. When an agent is unspecified — the default in most LLM interactions — the model defaults to abstract-executor assumptions from the pre-training distribution. The plan assumes infinite permissions, no physical constraints, and continuous execution because no specific agent's limitations have been loaded into the probability landscape.

5. Domain Consequences

5.1 Narrative: AI Fiction Writing

From a literary perspective, the who→when→why dependency explains a persistent failure in AI-generated fiction: characters who speak with no consistent temporal or spatial location. AI fiction often produces dialogue that is emotionally coherent but physically impossible — a character in two places simultaneously, a conversation that takes "a moment" but spans pages of text, cause and effect relationships that make emotional sense but violate temporal sequence.

The surface symptom is "bad writing" or "continuity errors." The structural cause is who(5) and when(3): the

model does not anchor characters in specific bodies at specific times, and therefore cannot track spatial-temporal consistency across narrative scenes. Causal plot logic (why) and character action logic (how) inherit these gaps.

Current mitigation — larger context windows, explicit scene notes — addresses the symptom (memory) rather than the cause (temporal computation deficit).

5.2 Analytical: Real-World Problem Analysis

From an analytical perspective, the dependency structure has severe implications for LLM deployment in policy, strategy, and diagnostic domains.

Post-hoc rationalization (why failure) combined with temporal missequencing (when failure) produces analyses that are structurally persuasive and causally wrong. An LLM analyzing a market crash may produce a coherent explanation with correct-sounding causal chains and incorrect temporal ordering — placing causes after effects, or compressing months-long feedback loops into simultaneous events.

The model's fluency masks the error. Unlike factual hallucination, where a wrong date can sometimes be verified, causal-temporal errors produce explanations that are internally consistent but externally false — and the consistency makes them harder to detect. This is the most dangerous failure mode for high-stakes analysis: the output *feels* rigorous because all the reasoning steps are present, but the dependency failure at when(3) cascades into why(7) producing structurally flawed conclusions.

5.3 Legal and Ethical: Forensic Misattribution

From a legal perspective, the combination of who and when failures creates acute danger in forensic and investigative applications — not through misattributing one named human for another, but through a deeper category error.

Misattribution across ontological categories. The who failure is not confusion between specific named individuals. It is confusion between ontological categories: real person versus fictional character, human statement versus AI-generated text, biographical author versus literary persona. In forensic contexts, this manifests as the model treating a fictional scenario as evidence, attributing an AI-generated output to a human source, or conflating a narrator's voice with an author's biography. LLMs exhibit a mercurial top-level ontology — their category boundaries are unstable, shifting with prompt context rather than tracking fixed distinctions between real and fictional, human and AI [24]. The growing difficulty of distinguishing LLM-generated text from human-authored text compounds this problem [25]. These errors survive retrieval augmentation because they are not factual retrieval failures — they are failures to track what *kind* of entity produced what *kind* of statement.

Correlation as causation. The why failure mode — post-hoc rationalization — is particularly dangerous in forensic contexts where temporal sequence is legally determinative. If an AI investigator observes Event A followed by Event B, the model may generate a causal explanation (A caused B) based purely on temporal adjacency, ignoring the distinction between *post hoc* and *propter hoc* that human legal reasoning has spent centuries articulating [26].

The temporal sequencing gap. Legal causation requires precise temporal ordering. Contract breach must precede damages. Negligence must precede harm. When an LLM with when(3/10) analyzes case timelines, the risk is not random error but systematic misordering — events placed in the most narratively coherent sequence rather than the factually correct one.

The existing legal-AI literature has focused on hallucination risk in legal research [27]. We identify a deeper problem: even when facts are retrieved correctly, the dependency structure (when→why) produces causal narratives that are internally coherent and legally wrong. This error survives retrieval augmentation because the retrieved facts are correctly identified but incorrectly sequenced and causally linked.

6. Implications

The dependency structure has practical implications for LLM deployment:

7. Limitations

This paper proposes a taxonomy and dependency structure based on observed failure patterns and existing benchmark results. We have not conducted novel empirical studies to validate the dependency claims. The baseline scores (5/10, 3/10, etc.) are heuristic estimates derived from operational experience with frontier models and published benchmark data; they are not precise measurements.

The dependency structure (when→why, who→how) is hypothesized from failure mode correlation and causal reasoning about model architecture. Controlled experiments — manipulating temporal information and measuring causal reasoning outcomes, manipulating agent identity and measuring planning executability — are needed to confirm the directionality and strength of these dependencies.

The framework is diagnostic: it identifies failure patterns and their relationships. It does not propose mitigation. The distinction between corpus-implicit failures and post-training artifacts is conceptual and requires empirical validation.

References

[1] Huang, L., et al. (2025). A Comprehensive Survey of Hallucination in Large Language Models. arXiv:2510.06265.

[2] Ji, Z., et al. (2024). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12).

[3] Farquhar, S., et al. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630.

[4] Sharma, M., et al. (2023). Entity-Level Hallucination in Large Language Models. *EMNLP 2023*.

[5] Bai, Y., et al. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.

[6] Perez, E., et al. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. arXiv:2212.09251.

[7] Desai, J. (2026). The Sycophancy Problem in Large Language Models. Whitepaper.

[8] Sharma, M., et al. (2025). Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence. *Science*.

[9] Gunjal, A., et al. (2024). Spatial Reasoning in Large Language Models: A Survey. arXiv:2405.02125.

[10] Pan, L., et al. (2024). Automatically Correcting Large Language Models: Surveying the Landscape of Diverse Self-Correction Strategies. *Transactions of the ACL*, 12.

[11] Gao, T., et al. (2025). RAC: Efficient LLM Factuality Correction with Retrieval Augmentation. *Findings of EMNLP 2025*.

[12] Fatemi, B., et al. (2024). Test of Time: A Benchmark for Evaluating LLMs on Temporal Reasoning. *ICLR 2025*.

[13] Qin, L., et al. (2021). TimeDial: Temporal Commonsense Reasoning in Dialog. *ACL 2021*.

[14] Zhou, Y., et al. (2025). TimE: A Multi-level Benchmark for Temporal Reasoning of LLMs in Real-World Scenarios. *NeurIPS 2025*.

[15] Jin, Z., et al. (2024). Failure Modes of LLMs for Causal Reasoning on Narratives. arXiv:2410.23884.

[16] Kıcıman, E., et al. (2023). Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. arXiv:2305.00050.

[17] Lyu, Q., et al. (2025). CoT-Faithfulness Survey. PKU-PILLAR Group, GitHub.

[18] Valmeekam, K., et al. (2023). On the Planning Abilities of Large Language Models — A Critical Investigation. *NeurIPS 2023*.

[19] Liu, Y., et al. (2025). Why Reasoning Fails to Plan: A Planning-Centric Analysis. arXiv:2601.22311.

[20] Wang, Z., et al. (2025). SPOC: Safety-Aware Planning Under Partial Observability and Physical Constraints. arXiv:2505.

[21] Ouyang, L., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. *NeurIPS 2022*.

[22] Vaswani, A., et al. (2017). Attention Is All You Need. *NeurIPS 2017*.

[23] Wei, J., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *NeurIPS 2022*.

[24] Jansen, F., & Verheij, B. (2025). The Mercurial Top-Level Ontology of Large Language Models. *SAGE Publications*.

[25] Zhou, Y., et al. (2024). Authorship Attribution in the Era of LLMs: Problems, Methodologies, and Challenges. *Proceedings of ACM Web Science Conference 2025*. arXiv:2408.08946.

[26] Wagner, G. (2024). Causal AI — A VISOR for the Law of Torts. *University of Chicago Law Review*.

[27] Dahl, M., et al. (2023). Hallucinating Law: Legal Mistakes in Large Language Models. *Proceedings of ICAIL 2023*.