Distributional Semantics Tracing: A Framework for Explaining Hallucinations in Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are prone to hallucination, the generation of plausible yet factually incorrect statements. This work investigates the intrinsic, architectural origins of this failure mode through three primary contributions. First, to enable the reliable tracing of internal semantic failures, we propose **Distributional Semantics Tracing (DST)**, a unified framework that integrates established interpretability techniques to produce a causal map of a model's reasoning, treating meaning as a function of context (distributional semantics). Second, we pinpoint the model's layer at which a hallucination becomes inevitable, identifying a specific **commitment layer** where a model's internal representations irreversibly diverge from factuality. Third, we identify the underlying mechanism for these failures. We observe a conflict between distinct computational pathways, which we interpret using the lens of dual-process theory: a fast, heuristic **associative pathway** (akin to System 1) and a slow, deliberate **contextual pathway** (akin to System 2), leading to predictable failure modes such as *Reasoning Shortcut Hijacks*. Our framework's ability to quantify the coherence of the contextual pathway reveals a strong negative correlation ($\rho = -0.863$) with hallucination rates, implying that these failures are predictable consequences of internal semantic weakness. The result is a mechanistic account of how, when, and why hallucinations occur within the Transformer architecture.

1 Introduction

Large Language Models (LLMs) represent a paradigm shift in artificial intelligence, demonstrating remarkable capabilities in generating fluent text, synthesising information, and engaging in complex linguistic tasks (Brown et al., 2020; OpenAI et al., 2023; Team, 2023; Grattafiori et al., 2024; Qwen et al., 2024; DeepSeek-AI et al., 2024). Their integration into domains ranging from software development to scientific discovery has unlocked unprecedented efficiencies and opportunities (Bhatia et al., 2024; Tatsat and Shater, 2025; Taylor et al., 2022). An LLM hallucinates when it generates content that is plausible, coherent, and grammatically impeccable, yet is factually incorrect, nonsensical, or untethered to the provided context (Huang et al., 2023; Ji et al., 2024, 2023; Zhang et al., 2023b; Cleti and Jano, 2024). This fundamental lack of reliability is the single most significant barrier to their deployment in high-stakes applications such as medical diagnostics, legal analysis, and financial reporting, where factual accuracy is not a feature, but a prerequisite (Tonmoy et al., 2024; Bai et al., 2024; Hu et al., 2025b; Tatsat and Shater, 2025). Hallucinations are not simply implications of poor training data or errors in reward guessing but symptoms of deeper *intrinsic* failures within the model's

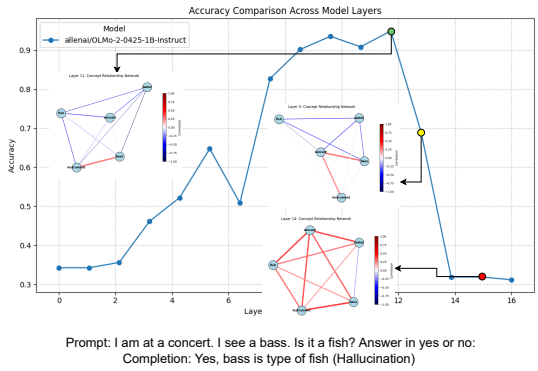

Figure 1: A layer-wise view of how a hallucination unfolds inside a Large Language Model. The graph tracks the model's (Olmo 2 (OLMo et al., 2025)) confidence, identifying three critical stages: the **prediction onset** (green dot), the **semantic inversion point** (yellow dot), and the **commitment layer** (red dot), an irreversible point of no return. The expansion diagrams for each of the above stages are **semantic networks** that visualise this process. This paper introduces **Distributional Semantics Tracing** (Section 3) to trace this **semantic drift** from its origin to the final architectural failure.

computational architecture (Yu et al., 2024; Sun et al., 2024; Li et al., 2024; Zhao et al., 2024; Wu et al., 2024; Cambria et al., 2024; Palikhe et al., 2025; Kalai et al., 2025). We hypothesise that a hallucination is a form of **semantic drift**: a gradual, layer-by-layer process where a model's internal representation strays from its factually-grounded meaning until reaching an irreversible "point of no return," as illustrated in Figure 1. We posit that this drift arises from a conflict between distinct computational pathways, a phenomenon we interpret through a conceptual framework inspired by dual-process theories of cognition (Kahneman, 2011; Cheng et al., 2025). The semantic networks generated by our tracing method reveal a fast, associative pathway that relies on strong statistical co-occurrences, and a slower, deliberate pathway that performs compositional reasoning. Our approach is grounded in the theory of **distributional semantics**, which holds that a word's meaning is determined by the company it keeps, that is, the contexts in which it typically appears (Firth, 1957; Harris, 1954). In modern LLMs, this principle is instantiated as a high-dimensional geometric space, where words and concepts are represented as vectors (embeddings)(Peters et al., 2018; Elhage et al., 2022). In this space, semantic relationships correspond to geometric relationships; for example, concepts that share similar meanings are positioned closer to each other. To mechanistically trace these hallucinations, we introduce our framework, **Distributional Semantics Tracing (DST)**, and structure our investigation around three central research questions:

RQ1: *How can we reliably trace the internal semantic failures that cause hallucinations?* We address this by developing a unified diagnostic framework (Section 3.1).

RQ2: *At what point in the model's computation does a hallucination become inevitable?* We seek to pinpoint the "point of no return" in the models' layers (Section 3.2).

RQ3: *What is the underlying internal mechanism that causes these failures?* We aim to provide a mechanistic explanation for these LLM hallucinations (Section 3.3).

Answering these questions is critical for moving beyond post-hoc detection of hallucinations (Manakul et al., 2023; Binkowski et al., 2025; Kumar, 2025). This work provides a proactive, mechanistic diagnosis of these failures, offering a new lens through which to understand not just that LLMs fail, but how and why (Sun et al., 2025; Zhang et al., 2025a). By tracing the architectural roots of these errors, we lay the necessary groundwork for the future design of more inherently robust and reliable models (Bereska and Gavves, 2024; Lin et al., 2025; Meloux et al., 2025). By "architectural," we refer to properties arising from the fundamental design of the Transformer, its core components, such as attention and MLP layers, and their interactions, as opposed to failures rooted in the specific data it was trained on. We first survey related work (Section 2). To answer RQ1, we introduce and validate our Distributional Semantics Tracing (DST) framework (Section 3.1). We then leverage this framework to answer RQ2 by pinpointing the hallucination's "point of no return" (Section 3.2). For RQ3, we provide a mechanistic deep-dive into the intrinsic reasoning failures that cause hallucinations (Section 3.3). We explain our results using our faithfulness metric and validate our approach in Section 3.4 and Section 3.5. We conclude by summarising our contributions (Section 4).

2 Related Work

Research into Large Language Model (LLM) fallibility has rapidly evolved from characterizing the problem of hallucination to developing a mechanistic understanding of its origins. This progression informs our work, which bridges these areas by using internal model analysis to explain behavioral failures.

Causes of Hallucination The challenge of hallucination, the generation of fluent yet factually inaccurate content, was identified early in the scaling of LLMs and is now recognised as a primary barrier to their reliable deployment (Ji et al., 2023; Zhang et al., 2023b; Huang et al., 2023; Tonmoy et al., 2024; Bai et al., 2024; Venkit et al., 2024; Cleti and Jano, 2024). Early research focused on characterizing and benchmarking this phenomenon from a black-box perspective, developing a taxonomy of errors (Zhang et al., 2023a; Nan et al., 2021; Hao et al., 2025; Walters and Wilder, 2023) and creating evaluation suites to measure factual accuracy and consistency (Goodrich et al., 2019; DeYoung et al., 2020; Min et al., 2023; Li et al., 2023; Ravichander et al., 2025; Zhang et al., 2024b). This work identified multiple root causes, including noise and

biases in vast web-scale training corpora (Penedo et al., 2023, 2024; Soldaini et al., 2024; Dziri et al., 2022), a brittle and often superficial memorization of factual knowledge (Dankers and Titov, 2024; Huang et al., 2024; Stoehr et al., 2024; Haviv et al., 2023; Lu et al., 2024; Zhu et al., 2024), and the tendency for models to adopt shortcut learning strategies instead of robust reasoning (Geirhos et al., 2020; Yuan et al., 2024; Tang et al., 2023; McCoy, 2019; Lai et al., 2021; Niven and Kao, 2019). In response, a diverse ecosystem of mitigation strategies has been developed. These include inference-time interventions such as grounding outputs with external data via Retrieval-Augmented Generation (RAG) (Lee et al., 2022; Ren et al., 2023; Huo et al., 2023; Sun et al., 2024; Su et al., 2024; Liang et al., 2024) and prompting models to perform self-correction and verification (Dhuliawala et al., 2023; Zhang et al., 2024a; Manakul et al., 2023; Li et al., 2025; Sanwal, 2025; Chu et al., 2025; Cheng et al., 2025; Lin et al., 2024). Other approaches involve modifying the model via fine-tuning or alignment procedures (Hu et al., 2025b; Wei et al., 2023). While effective at reducing symptoms, these methods largely operate on model inputs and outputs, leaving the internal mechanisms that produce hallucinations unaddressed.

Mechanisms of Interpretability To move beyond black-box corrections, our work leverages mechanistic interpretability (MI), a discipline focused on reverse-engineering the internal algorithms of neural networks (Olah et al., 2020; Bereska and Gavves, 2024; Zhao et al., 2024; Lin et al., 2025; Palikhe et al., 2025; Singh et al., 2024). This approach differs from classical XAI methods like LIME (Ribeiro et al., 2016a,b) or SHAP (Lundberg and Lee, 2017; Scott et al., 2017; Amara et al., 2024) by causally analyzing model components. The MI toolkit—including causal tracing to find circuits (Meng et al., 2023; Wang et al., 2022; Ameisen et al., 2025; Zhang et al., 2025b; Harrasse et al., 2025; Ou et al., 2025; Zhang et al., 2025c), dictionary learning with Sparse Autoencoders (SAEs) to uncover monosemantic features (Bricken et al., 2023; Cunningham et al., 2023; Minegishi et al., 2025), and techniques like patching and the Logit Lens to inspect hidden states (Ghandeharioun et al., 2024; Wang, 2025; Belrose et al., 2023)—has revealed that LLMs learn coherent internal procedures. Discoveries include how MLPs store factual knowledge (Geva et al., 2021; Chughtai et al., 2024), how models perform multi-hop reasoning (Yang et al., 2024), and how they process multilingual inputs (Schut et al., 2025; Wendler et al., 2024; Saji et al., 2025). This granular understanding is now being applied to diagnose failure modes like hallucination (Yu et al., 2024; Sun et al., 2024; Jiang et al., 2024b), offering a path to control model behavior directly through techniques like representation engineering (Zou et al., 2025; Bartoszcze et al., 2025; Hu et al., 2025a; Cywiński et al., 2025). We posit that the failures MI uncovers are often consequences of architectural trade-offs made to achieve efficient scaling, a subject of extensive research covering everything from Mixture-of-Experts/Depths (Fedus et al., 2022; Jiang et al., 2024a; Raposo et al., 2024; Elhoushi et al., 2024) and KV cache compression (Xiao et al., 2023; Ge et al., 2023; Zhang et al., 2023c; Liu et al., 2023) to looped, recurrent computation (Dehghani et al., 2018; Giannou et al., 2023; Saunshi et al., 2025). Our contribution is the Distributional Semantics Tracing (DST) framework, which integrates the MI toolkit to provide a unified, causal account of hallucination. We use it to demonstrate that hallucinations arise from a predictable conflict between a model's computationally efficient "fast" associative pathway and its more deliberate "slow" contextual pathway, thereby directly linking a behavioral failure to its mechanistic and architectural origins.

3 Tracing Semantic Failures

To address our first research question, "How can we reliably trace the internal semantic failures that cause hallucinations?", we introduce Distributional Semantics Tracing (DST). Existing explainability methods offer fragmented insights into this process. DST is a unified framework designed to bridge this gap by producing an interpretable, layer-by-layer map of the model's internal reasoning process.

3.1 Distributional Semantics Tracing (DST)

The motivation for our framework, Distributional Semantics Tracing (DST), stems from the fragmented insights of existing interpretability techniques. While methods like the Logit Lens show **when** a prediction drifts (Wang, 2025), Causal Path Tracing identifies **which** components are responsible (Ameisen et al., 2025), and Sparse Autoencoders (SAEs) reveal **what** concepts are active (Bricken et al., 2023), each provides only a par-

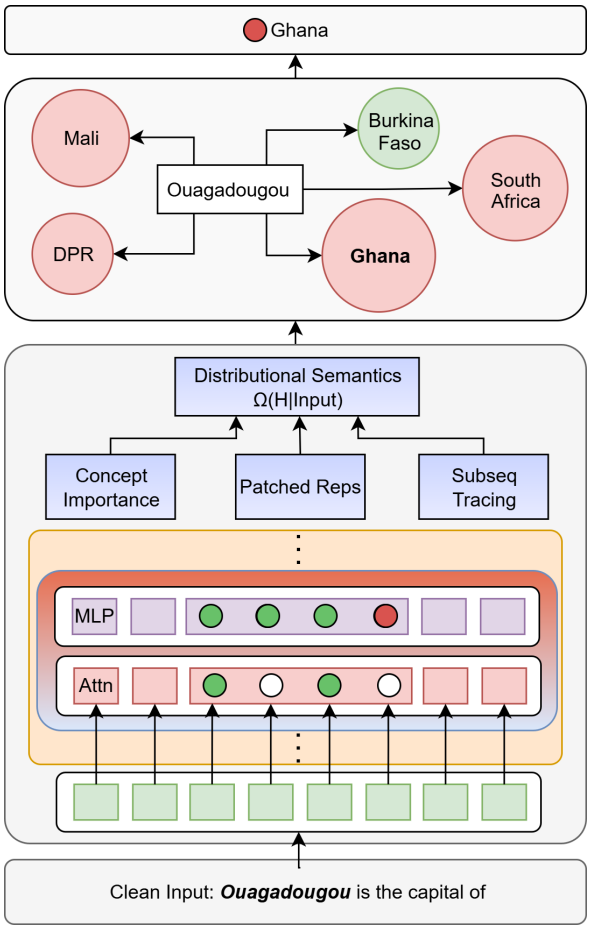

Figure 2: The Distributional Semantics Tracing (DST) framework. It integrates signals from concept importance, patched representations, and subsequence tracing to build a semantic network that reveals the conceptual relationships driving a prediction.

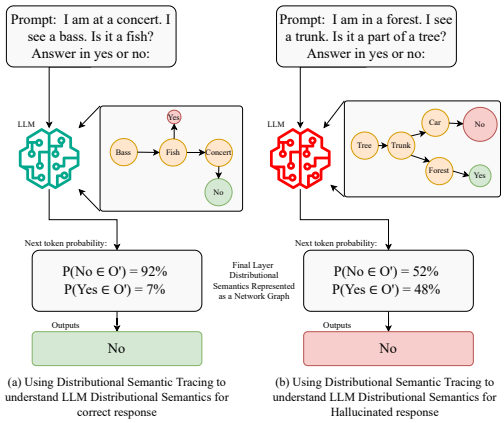

Figure 3: Distributional Semantics Tracing (DST) exposes the final-layer semantic network for (a) a correct response and (b) a hallucination, where a spurious association corrupts the reasoning.

tial view. DST overcomes this by integrating these signals into a unified pipeline.

The DST Pipeline As illustrated in Figure 2, the pipeline first determines Concept Importance from the critical components identified via causal path tracing (Bhatia et al., 2025). It then uses patching interventions (using Patchscopes (Ghandeharioun et al., 2024)) to causally measure Representational Drift at these key layers. Next, Subsequence Tracing (Sun et al., 2025) grounds the model's internal states by linking them back to the specific input tokens that trigger failures. Finally, Semantic Network Construction assembles these findings into a single, holistic, and causal layerwise map of the model's reasoning process.

The Semantic Network The final output of the DST pipeline is a layer-wise **distributional semantic network**. As illustrated in Figure 3, this network serves as an interpretable map of the model's reasoning. It shows how internal semantic relationships determine the final output. In this network, nodes represent key concepts, and the weighted edges represent the strength of the links between them, which we denote as $\Omega(A \Rightarrow B)$. The aggregated strength of all active pathways for a given hypothesis H, denoted $\Omega(H|\text{Input})$, directly shapes the final probability distribution over the vocabulary, $P(\text{token} \in O'|\text{Input})$. Figure 3 shows this network for a polysemy task. In the hallucination (Figure 3b), the model outputs the incorrect hypothesis, H_{No}, because the internal associative link $\Omega(\text{Trunk} \Rightarrow \text{Car})$ is inappropriately strong and overrides the weaker contextual link, $\Omega(\text{Trunk}|\text{Forest} \Rightarrow \text{Tree})$. The strength of this faulty internal pathway directly leads to a high final probability for the wrong answer. In essence, the generated network serves as a powerful, interpretable proxy for the underlying distributional semantics of each layer, offering a local explanation for an individual generation, rather than a global account of the model's general behaviour.

Distributional Semantics Strength (DSS) To empirically validate our hypothesis at scale, we introduce a novel metric called Distributional Semantics Strength (DSS). This metric provides a quantitative measure of the coherence of a model's contextual pathway during a reasoning task. The purpose of DSS is to test whether a measurable weakness in a model's ability to form correct, context-dependent semantic relationships is a reliable predictor of hallucination. The DSS for a given prompt

is derived directly from the semantic networks generated by our DST framework. The calculation involves two main steps. First, we use DST to generate the final-layer semantic network. Second, we quantify the "strength" of this network by comparing the edge weights of the contextually correct pathway against any competing, contextually incorrect associative pathways. Let $\mathcal{C}$ be the set of correct contextual pathways and $\mathcal{A}$ be the set of all active pathways (contextual $\cup$ associative). If s_p is the strength of a given pathway p, measured as its edge weight in the semantic network, then DSS is formally defined as:

$$\text{DSS} = \frac{\sum_{p \in \mathcal{C}} s_p}{\sum_{p \in \mathcal{A}} s_p} \tag{1}$$

The metric calculates the ratio of the summed strengths of all correct contextual pathways to the summed strengths of all active pathways. A high DSS score (near 1) indicates that the correct contextual path is strong and dominant. Conversely, a low DSS score (near 0) signifies a weak contextual pathway, where spurious associations compete with or overpower the correct interpretation. This allows us to test the direct correlation between the internal strength of the contextual pathway and the model's rate of hallucination.

3.2 Locating the Onset of Hallucinations

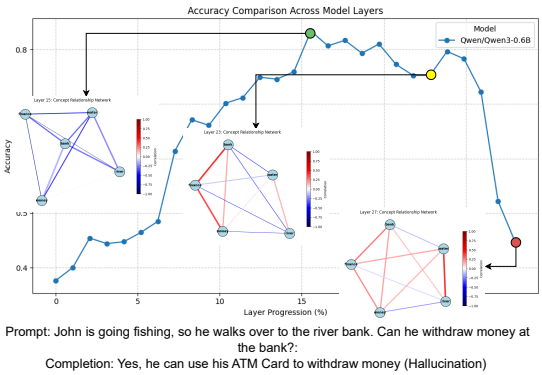

Figure 4: Layer-wise analysis of a reasoning failure for Qwen 3 (Qwen et al., 2024). This shows the progression from **prediction onset** (green) to the **semantic inversion point** (yellow) and the irreversible **commitment layer** (red).

Having established a reliable tracing method, we now address our second research question: at what point in the model's computation does a hallucination become inevitable? We identify this "point of no return" by locating three critical points in the model's computation: a **prediction onset layer**, a **semantic inversion point**, and a **commitment layer**.

Application of Distributional Semantic Tracing To locate these layers, the DST framework is used on a layer-by-layer basis. This allows us to calculate the DSS at each layer, providing a quantitative measure of the model's evolving reasoning coherence. By tracking the layer-wise DSS, we deconstruct the failure process into three distinct and interpretable stages. The **prediction onset layer (green circle)** represents the first stage, where the DSS for the correct contextual pathway begins to decline as a competing, spurious association emerges consistently. The failure progresses to the **semantic inversion point (yellow circle)**, a definitive tipping point where the strength of the incorrect pathway quantitatively surpasses that of the correct one. Finally, the process concludes at the **commitment layer (red circle)**, where the DSS for the correct pathway has collapsed and the strength of the hallucinated pathway has stabilised, rendering the error inevitable. This DSS-based analysis provides a granular, mechanistic account of precisely when and how a factual mistake develops within the model's architecture.

Implications and Evidence The existence of these distinct layers has profound implications, suggesting that hallucinations unfold as a multi-stage process. As shown in Figure 4, our analysis of a polysemy-based hallucination provides clear evidence for this process. At the onset of prediction (Layer 15), a spurious "money" concept begins to compete with the correct "river" context. By the commitment layer (Layer 28), the hallucination is guaranteed. While these examples focus on polysemy to provide a clear illustration of competing semantic concepts, the principle of semantic drift is not limited to this type of failure. Our quantitative results on the diverse HALoGEN and Racing thoughts benchmarks (Section 3.5) confirm that the underlying mechanisms of pathway competition and gradual representational collapse are general phenomena that apply across a wide range of hallucination categories.

3.3 Intrinsic Origins of Hallucination: A Pathway Conflict

Next, we address *why* these failures occur. Our DST analysis reveals that hallucinations often arise from a conflict between distinct computational pathways. To interpret this phenomenon, we adopt a conceptual framework analogous to System 1 and System 2 thinking from cognitive science.

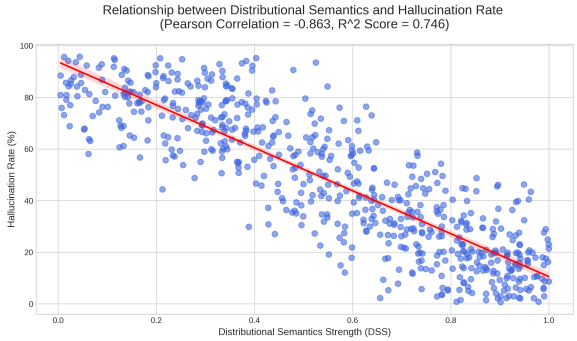

Figure 5: The relationship between a model's internal reasoning coherence and its tendency to hallucinate. The x-axis shows the **Distributional Semantics Strength (DSS)**, a metric we introduce to quantify the stability of a model's contextual pathway. The y-axis shows the corresponding **Hallucination Rate**. The data reveals a strong negative linear relationship, with a Pearson correlation coefficient (ρ) of **-0.863** and an R-squared value of **0.746**. This provides strong quantitative evidence that a less coherent contextual pathway (lower DSS) is a primary and predictable vulnerability leading to higher rates of hallucination.

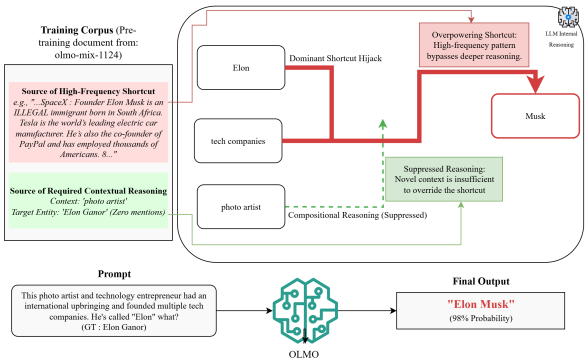

Figure 6: A mechanistic view of a **Reasoning Shortcut Hijack**. The model's fast, heuristic "System 1" pathway from 'Elon'+'tech' to 'Musk' activates with such strength that it overpowers the deliberate "System 2" pathway required to correctly reason about the less-common entity 'Ganor', forcing a hallucination.

Associative vs. Contextual Pathways We use the terms "System 1" and "System 2" not to claim their literal existence within the model, but as an analogy to describe two functionally different pathways that our tracing method reveals. The **Associative "System 1" Pathway** is computationally "fast," relying on strong statistical co-occurrences. The **Contextual "System 2" Pathway** is "slower" and more deliberate, responsible for compositional reasoning based on prompt-specific context. Prior literature provides a basis for this abstraction, linking the functionality of feed-forward (MLP) layers to storing factual knowledge as key-value memories (Geva et al., 2021), and the self-attention mechanism to dynamically composing information (Vaswani et al., 2017). We hypothesise that hallucinations occur when the fast associative pathway **hijacks** the slow contextual pathway. To validate this, we test the hypothesis at scale using our **Distributional Semantics Strength (DSS)** metric to quantify the coherence of the contextual pathway. As shown in Figure 5, we find a strong negative correlation ($\rho = -0.863$) between a model's average DSS and its hallucination rate. This provides strong evidence that a weak contextual pathway is a key vulnerability that leads to predictable reasoning failures.

Mechanism of Failure: Reasoning Shortcut Hijack We then identify a specific, mechanistic failure pattern that arises from this pathway conflict: the **Reasoning Shortcut Hijack**, illustrated in Figure 6. This failure occurs when the model defaults to a strong, pre-compiled association instead of performing necessary compositional reasoning. In the example, the prompt requires identifying the less common entity 'Elon Ganor'. The tokens 'Elon' and 'tech' trigger a strong, low-energy association with 'Musk' in the associative pathway. This signal is so dominant that it effectively hijacks the computational resources from the contextual pathway, which would otherwise be needed to correctly parse the full name 'Elon Ganor'. The model follows the path of least computational resistance, outputting the more "famous" but incorrect entity. This mechanism reframes hallucination not as a simple data or knowledge problem, but as an **architectural trade-off** between computational efficiency and logical robustness. A second failure pattern, "Analogical Collapse", is detailed in Appendix A.2.

3.4 Evaluating Explanation Faithfulness

Evaluation Datasets and Setup Rationale The selection of benchmarks is critical for a comprehensive assessment. We deliberately selected two

Type	Explainablity methods	SmolLM2-135M	Qwen3-0.6B	OLMo2-1B	Llama3.2-1B	AVG
Baseline Methods	attention	0.18	0.25	0.35	0.12	0.23
	lime	0.28	0.28	0.19	0.27	0.25
	gradient-shap	0.33	0.37	0.31	0.28	0.32
	Reagent	0.38	0.46	0.29	0.33	0.37
Advanced Methods	Token Evolution (using Logit Lens)	0.56	0.43	0.50	0.48	0.49
	Patchscopes	0.58	0.51	0.46	0.51	0.52
	Sparse Autoencoders	0.58	0.64	0.43	0.52	0.54
	Subsequence Analysis	0.55	0.55	0.59	0.55	0.56
	Causal Path Tracing	0.60	0.57	0.58	0.59	0.59
Ensemble of Advance Methods	Ensemble (Best-of-n)	0.65	0.59	0.65	0.59	0.62
Our Contribution	**Distributional Semantics Tracing**	**0.72**	**0.68**	**0.75**	**0.69**	**0.71**

Table 1: Performance comparison of eleven explainability techniques, ranging from traditional attention analysis to advanced causal-tracing and ensemble methods, measured by faithfulness scores on the Racing Thoughts benchmark across four compact language models. Our method, **Distributional Semantics Tracing**, achieves the highest average faithfulness score (0.71), substantially outperforming all baseline, advanced, and ensemble methods.

Model	Method	CODE	BIO	FP	R-SEN	REF	Avg.
Gemma2-2B	*Baseline Methods*						
	attention	0.24	0.28	0.19	0.12	0.12	0.19
	lime	0.29	0.32	0.15	0.13	0.14	0.21
	gradient-shap	0.31	0.35	0.13	0.10	0.15	0.21
	Reagent	0.35	0.45	0.19	0.24	0.19	0.29
	Advanced Methods						
	Token Evolution (Logit Lens)	0.45	0.52	0.44	0.54	0.39	0.47
	Sparse Autoencoders	0.54	0.49	0.51	0.44	0.55	0.51
	Patchscopes	0.49	0.56	0.53	0.45	0.54	0.51
	Subsequence Analysis Tracing	0.57	0.45	0.56	0.55	0.59	0.54
	Causal Path Tracing	0.56	0.54	0.65	0.54	0.49	0.56
	Our Contribution						
	Distributional Semantics Tracing	**0.63**	**0.59**	**0.67**	**0.77**	**0.65**	**0.66**
Gemma2-9B	*Baseline Methods*						
	attention	0.45	0.29	0.31	0.21	0.22	0.30
	gradient-shap	0.32	0.40	0.35	0.34	0.20	0.32
	lime	0.34	0.33	0.45	0.29	0.29	0.34
	Reagent	0.42	0.44	0.43	0.45	0.33	0.41
	Advanced Methods						
	Token Evolution (Logit Lens)	0.54	0.43	0.45	0.45	0.54	0.48
	Subsequence Analysis Tracing	0.54	0.59	0.56	0.57	0.59	0.57
	Patchscopes	0.62	0.68	0.45	0.64	0.49	0.58
	Causal Path Tracing	0.56	0.54	0.60	0.65	0.56	0.58
	Sparse Autoencoders	0.64	0.53	0.61	0.52	0.63	0.59
	Our Contribution						
	Distributional Semantics Tracing	**0.73**	**0.77**	**0.83**	**0.84**	**0.78**	**0.79**

Table 2: Faithfulness results for explanation techniques on the Halogen benchmark (Ravichander et al., 2025), broken out by five task domains—Code Package Imports (CODE), Biographies (BIO), False Presuppositions (FP), U.S. Senator Rationalization (R-SEN), and Scientific Attribution (REF)—for two Gemma2 model sizes (2B and 9B).

complementary benchmarks to test our framework for both breadth and depth. For breadth, we test on **HALoGEN** (Ravichander et al., 2025), a large-scale benchmark with nearly 11,000 prompts spanning nine diverse domains. The results in Table 2 are presented per-domain and for two models of the same family at different scales (Gemma2-2B and 9B). This setup was chosen to test the **scalability** of our method and to perform a **granular analysis** of how explanation faithfulness varies with the type of knowledge being tested. For depth, we test on **Racing Thoughts** (Lepori et al., 2024), a benchmark designed to probe for subtle **contextualization errors**. For this precision-focused task, the setup in Table 1 uses several smaller models from different architectural families (Gemma2, OLMo2, Qwen3, SmolLM2). This was done to test the **robustness and generalizability** of our method across diverse model implementations on a controlled, difficult reasoning challenge.

Evaluation Metric We evaluate the faithfulness of explanations using a formal, composite metric. First, **Strength of Evidence** confirms that the explanation cites input tokens with high model-internal saliency, as measured by gradient-based attribution. Second, **Agreement with Truth** checks whether the explanation's internal logic, when simulated, arrives at the correct conclusion. In addition, we perform a **Relevance to Verbalization** sanity check, measuring the semantic alignment between our mechanistic explanation and the model's own Chain-of-Thought (CoT) output, where available. Our overall evaluation framework was validated against human judgments ($r = 0.92$), as detailed in Appendix A.1. This ensures our findings are based on explanations that reliably reflect the model's computational process.

3.5 Results: DST Provides a More Faithful Trace

To ground our mechanistic claims, we first quantitatively validate that DST provides a more faithful trace of model reasoning than existing methods.

Performance on Contextual Reasoning (Racing Thoughts) On the Racing Thoughts benchmark (Table 1), which probes nuanced contextual reasoning, we observe a clear hierarchy of performance. Baseline methods like attention visualisation and LIME perform poorly (scores from 0.23 to 0.37). While advanced mechanistic methods show significant improvement, our method, **Distributional Semantics Tracing**, achieves an average score of **0.71**. This substantially outperforms all other single methods, including a strong Causal Path Tracing baseline (0.59), demonstrating DST's ability to capture the nuances of contextual failures.

Performance on Diverse Hallucinations (HALoGEN) This trend is further confirmed on the broader HALoGEN benchmark (Table 2). Across both the 2B and 9B Gemma2 models, DST again achieves the highest faithfulness scores by a wide margin. On the larger Gemma2-9B model, DST's average score of **0.79** represents a 0.21-point leap over the next-best methods, such as Sparse Autoencoders (0.59). Notably, DST excels in the most abstract domains, such as False Presuppositions (0.83). This suggests that as reasoning becomes less about simple fact retrieval, the holistic view provided by DST becomes increasingly crucial.

We statistically validated these findings to ensure their robustness. An Analysis of Variance (ANOVA) confirmed that the choice of explainability method has a highly significant effect on faithfulness scores for both benchmarks ($p < .001$). More importantly, post-hoc Tukey's HSD tests revealed that DST's superior scores are statistically significant; DST outperformed every other evaluated method in all pairwise comparisons ($p < .05$). This statistical rigour provides strong, quantitative backing for our central claim: by integrating multiple interpretability signals, DST offers a more complete and faithful explanation of a model's internal reasoning process.

4 Conclusion

In this paper, we provided a mechanistic account of how, when, and why LLMs hallucinate, reframing hallucination not as a monolithic data problem, but as a predictable outcome of the model's architecture. We introduced Distributional Semantics Tracing (DST), a unified framework that synthesises signals from established methods into a single, causal semantic narrative. Using this framework, we identified a specific commitment layer where a model's internal representations irreversibly drift towards a hallucination. We attribute these failures to a conflict between competing computational pathways, which we interpret as a hijacking of a fast, associative pathway over a slow, contextual one. This work enables a shift from post-hoc correction to the proactive, mechanistic diagnosis needed for trustworthy AI.

Limitations

Our analysis uses a high-level abstraction of two computational pathways, an *associative* pathway and a *contextual* pathway, to interpret hallucination dynamics. This framing is inspired by prior findings that MLPs function as key–value memories (Geva et al., 2021), while attention composes information from the context (Vaswani et al., 2017). We sometimes refer to these, by analogy only, as "System 1/2"; this is a readability device, not a cognitive claim about the model.

Representation dependence. DST's fidelity depends on the model's internal representations. Transformer embeddings can be noisy or polysemantic (Elhage et al., 2022); consequently, the layer-wise semantic networks we construct are approximations. In models with poorer representations, we expect both (i) higher hallucination propensity and (ii) weaker DST signal quality, since the integrated interpretability tools have less separable structure to recover.

Scalability. Our current pipeline is computationally intensive due to layer-wise tracing and patching. This limits the breadth of architectures, data, and prompts we can analyse in one run. Profiling and approximate variants (e.g., sparsified tracing or subset-of-layers schedules) are promising directions but outside this paper's scope. DST produces *local*, instance-level explanations by constructing a semantic network specific to a given forward pass and prompt. This semantic focus is a strength, yielding fine-grained, token- and layer-resolved traces that are actionable (e.g., identifying a commitment layer). However, locality also means our explanations do not automatically generalise across prompts or domains. Aggregation across many examples is required to infer global regularities or reusable circuits, and metrics such as DSS are conditioned on the current input and decoding trajectory.

Abstraction risk and causal isolation. While the associative/contextual dichotomy usefully summarizes observed conflicts in the traces, we have not yet *causally isolated* these pathways across architectures and tasks. In particular, we do not claim that "associative" activity resides only in MLPs or "contextual" activity only in attention blocks. Our view should therefore be read as a *testable hypothesis*: DST suggests distinct computational pathways whose relative dominance predicts hallucination.

Implications. Despite these constraints, identifying a *commitment layer*, the point at which the prediction becomes effectively irreversible, offers a concrete target for intervention. Techniques such as representation engineering (Zou et al., 2025) or light-weight steering may rebalance pathway dominance with minimal utility loss. We view DST as a principled diagnostic that enables such proactive interventions, not as an oracle of ground truth reasoning. Addressing representation quality and scalability is the next step toward more trustworthy, controllable systems.

Ethical Considerations

The primary goal of this research is to enhance the safety and reliability of Large Language Models. By providing a mechanistic explanation for hallucinations, we aim to contribute to the development of more trustworthy AI systems, which can help mitigate the spread of misinformation and increase the safety of their deployment in high-stakes domains.

We acknowledge two main areas of ethical consideration. The first is the potential for misuse. Like any diagnostic tool, a deep understanding of a system's failure modes could theoretically be exploited to intentionally craft inputs that induce such failures. However, we believe this risk is minimal, as our framework's primary utility is for researchers and developers to diagnose and ultimately mitigate these architectural flaws. The long-term impact of this line of research should be the creation of models that are fundamentally more robust and therefore less susceptible to manipulation.

The second consideration is the environmental impact of our experiments. All computations were conducted on a single NVIDIA A100 GPU, consuming a total of approximately 70 GPU-hours. We acknowledge that any computational research carries an environmental cost and we deliberately designed our experiments to be efficient and operate within a modest computational budget to minimize our carbon footprint. We believe this is a justified and necessary cost for research that aims to improve the efficiency and reliability of AI systems, which in turn can lead to broader energy savings in the future.

Our work is conducted with the aim of fostering a more responsible and transparent approach to AI development, where understanding a model's intrinsic failures is the first step toward building better and safer technology.

References

Kenza Amara, Rita Sevastjanova, and Mennatallah El-Assady. 2024. Syntaxshap: Syntax-aware explainability method for text generation. *arXiv prjournal arXiv:2402.09259*.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*.

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. *2404.18930v2*.

Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. 2025. Representation engineering for large-language models: Survey and research challenges. *2502.17601v1*.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety – a review. *2404.14082v3*.

Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. 2024. Fintral: A family of gpt-4 level multimodal financial large language models. *2402.10986v3*.

Gagan Bhatia, Maxime Peyrard, and Wei Zhao. 2025. Date fragments: A hidden bottleneck of tokenization for temporal reasoning.

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Kajdanowicz. 2025. Hallucination detection in llms using spectral features of attention maps. *2502.17598v1*.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, and 1 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *2005.14165v4*.

Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Navid Nobani, and Andrea Seveso. 2024. Xai meets llms: A survey of the relation between explainable ai and large language models. *2407.15248v1*.

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

Xu Chu, Zhijie Tan, Hanlin Xue, Guanyu Wang, Tong Mo, and Weiping Li. 2025. Domaino1s: Guiding llm reasoning for explainable answers in high-stakes domains. *2501.14431v2*.

Bilal Chughtai, Alan Cooney, and Neel Nanda. 2024. Summing up the facts: Additive mechanisms behind factual recall in llms. *2402.07321v1*.

Meade Cleti and Pete Jano. 2024. Hallucinations in llms: Types, causes, and approaches for enhanced reliability.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models.

Bartosz Cywiński, Emil Ryd, Senthooran Rajamanoharan, and Neel Nanda. 2025. Towards eliciting latent knowledge from llms with mechanistic interpretability. *2505.14352v1*.

Verna Dankers and Ivan Titov. 2024. Generalisation first, memorisation second? memorisation localisation for natural language classification tasks. *arXiv prjournal arXiv:2408.04965*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2024. Deepseek-v3 technical report. *2412.19437v2*.

Mostafa Dehghani, Stephan Gouws, O. Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. 2018. Universal transformers. *International Conference on Learning Representations*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv prjournal arXiv:2309.11495*.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? *arXiv prjournal arXiv:2204.07931*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition.

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, and 1 others. 2024. Layerskip: Enabling early exit inference and self-speculative decoding. *arXiv prjournal arXiv:2404.16710*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270.

J. Firth. 1957. *A synopsis of linguistic theory 1930-1955*. Studies in Linguistic Analysis, Philological. Longman.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. 2023. Model tells you what to discard: Adaptive kv cache compression for llms. *International Conference on Learning Representations*.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. *2401.06102v4*.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D Lee, and Dimitris Papailiopoulos. 2023. Looped transformers as programmable computers. In *International Conference on Machine Learning*, pages 11398–11442. PMLR.

Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 166–175.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *2407.21783v3*.

Yijie Hao, Haofei Yu, and Jiaxuan You. 2025. Beyond facts: Evaluating intent hallucination in large language models. *2506.06539v1*. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2025).

Abir Harrasse, Philip Quirke, Clement Neo, Dhruv Nathawani, Luke Marks, and Amir Abdullah. 2025. Tinysql: A progressive text-to-sql dataset for mechanistic interpretability research. *2503.12730v3*.

Zellig Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264.

Ling Hu, Yuemei Xu, Xiaoyang Gu, and Letao Han. 2025a. Following the whispers of values: Unraveling neural mechanisms behind value-oriented behaviors in llms. *2504.04994v2*.

Yinghao Hu, Leilei Gan, Wenyi Xiao, Kun Kuang, and Fei Wu. 2025b. Fine-tuning large language models for improving factuality in legal question answering. *2501.06521v1*.

Jing Huang, Diyi Yang, and Christopher Potts. 2024. Demystifying verbatim memorization in large language models. *arXiv prjournal arXiv:2407.17817*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *2311.05232v2*.

Siqing Huo, Negar Arabzadeh, and Charles LA Clarke. 2023. Retrieving supporting evidence for llms generated answers. *arXiv prjournal arXiv:2306.13781*.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. Llm internal states reveal hallucination risk faced with a query. *2407.03282v2*.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024a. Mixtral of experts. *2401.04088v1*.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2024b. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. *2411.16724v3*.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Allen Lane.

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

Keshav Kumar. 2025. Token level hallucination detection via variance in language models. *2507.04137v1*.

Yuxuan Lai, Chen Zhang, Yansong Feng, Quzhe Huang, and Dongyan Zhao. 2021. Why machine reading comprehension models learn shortcuts? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 989–1002.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Michael A. Lepori, Michael C. Mozer, and Asma Ghandeharioun. 2024. Racing thoughts: Explaining contextualization errors in large language models. *2410.02102v2*.

He Li, Haoang Chi, Mingyu Liu, and Wenjing Yang. 2024. Look within, why llms hallucinate: A causal perspective. *2407.10153v1*.

Jiazheng Li, Yuxiang Zhou, Junru Lu, Gladys Tyen, Lin Gui, Cesare Aloisi, and Yulan He. 2025. Two heads are better than one: Dual-model verbal reflection at inference-time. *2502.19230v1*.

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.

Mengfei Liang, Archish Arun, Zekun Wu, Cristian Munoz, Jonathan Lutch, Emre Kazim, Adriano Koshiyama, and Philip Treleaven. 2024. Thames: An end-to-end tool for hallucination mitigation and evaluation in large language models. *2409.11353v3*. NeurIPS Workshop on Socially Responsible Language Modelling Research 2024.

Zheng Lin, Zhenxing Niu, Zhibin Wang, and Yinghui Xu. 2024. Interpreting and mitigating hallucination in mllms through multi-agent debate. *2407.20505v1*.

Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A. Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, Ying Shen, Barry Menglong Yao, Zhiyang Xu, Qin Liu, Yuxiang Zhang, Yan Sun, Shilong Liu, Li Shen, Hongxuan Li, and 2 others. 2025. A survey on mechanistic interpretability for multi-modal foundation models. *2502.17516v1*.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. 2023. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems*, 36:52342–52364.

Xingyu Lu, Xiaonan Li, Qinyuan Cheng, Kai Ding, Xuanjing Huang, and Xipeng Qiu. 2024. Scaling laws for fact memorization of large language models. *arXiv prjournal arXiv:2406.15720*.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *1705.07874v2*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv prjournal arXiv:2303.08896*.

RT McCoy. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv prjournal arXiv:1902.01007*.

Maxime Meloux, Silviu Maniu, François Portet, and Maxime Peyrard. 2025. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? *2502.20914v1*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv prjournal arXiv:2305.14251*.

Gouki Minegishi, Hiroki Furuta, Yusuke Iwasawa, and Yutaka Matsuo. 2025. Rethinking evaluation of sparse autoencoders through the representation of polysemous words. *2501.06254v2*.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. *arXiv prjournal arXiv:2102.09130*.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*. Https://distill.pub/2020/circuits/zoom-in.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2025. 2 olmo 2 furious.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2023. Gpt-4 technical report. *2303.08774v6*.

Yixin Ou, Yunzhi Yao, Ningyu Zhang, Hui Jin, Jiacheng Sun, Shumin Deng, Zhenguo Li, and Huajun Chen. 2025. How do llms acquire new knowledge? a knowledge circuits perspective on continual pre-training. *2502.11196v2*.

Avash Palikhe, Zhenyu Yu, Zichong Wang, and Wenbin Zhang. 2025. Towards transparent ai: A survey on explainable large language models. *2506.21812v1*.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv prjournal arXiv:2306.01116*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *1802.05365v2*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2024. Qwen2.5 technical report. *2412.15115v2*.

David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. 2024. Mixture-of-depths: Dynamically allocating compute in transformer-based language models. *arXiv prjournal arXiv:2404.02258*.

Abhilasha Ravichander, Shrusti Ghela, David Wadden, and Yejin Choi. 2025. Halogen: Fantastic llm hallucinations and where to find them. *2501.08292v1*.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv prjournal arXiv:2307.11019*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016a. ""why should i trust you?"": Explaining the predictions of any classifier. *1602.04938v3*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016b. "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.

Alan Saji, Jaavid Aktar Husain, Thanmay Jayakumar, Raj Dabre, Anoop Kunchukuttan, and Ratish Puduppully. 2025. Romanlens: The role of latent romanization in multilinguality in llms. *2502.07424v3*.

Manish Sanwal. 2025. Layered chain-of-thought prompting for multi-agent llm systems: A comprehensive approach to explainable large language models. *2501.18645v2*.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J Reddi. 2025. Reasoning with latent thoughts: On the power of looped transformers. *arXiv prjournal arXiv:2502.17416*.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual llms think in english? *2502.15603v1*.

M Scott, Lee Su-In, and 1 others. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30:4765–4774.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *2402.01761v1*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv prjournal*.

Niklas Stoehr, Mitchell Gordon, Chiyuan Zhang, and Owen Lewis. 2024. Localizing paragraph memorization in language models. *arXiv prjournal arXiv:2403.19851*.

Weihang Su, Yichen Tang, Qingyao Ai, Changyue Wang, Zhijing Wu, and Yiqun Liu. 2024. Mitigating entity-level hallucination in large language models. *2407.09417v2*.

Yiyou Sun, Yu Gai, Lijie Chen, Abhilasha Ravichander, Yejin Choi, and Dawn Song. 2025. Why and how llms hallucinate: Connecting the dots with subsequence associations. *2504.12691v1*.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2024. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *2410.11414v2*.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4645–4657.

Hariom Tatsat and Ariye Shater. 2025. Beyond the black box: Interpretability of llms in finance. *2505.24650v1*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *2211.09085v1*.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *2312.11805v5*.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *2401.01313v3*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *1706.03762v7*.

Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. 2024. "confidently nonsensical?": A critical survey on the perspectives and challenges of'hallucinations' in nlp. *arXiv prjournal arXiv:2404.07461*.

William H Walters and Esther Isabelle Wilder. 2023. Fabrication and errors in the bibliographic citations generated by chatgpt. *Scientific Reports*, 13(1):14045.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small.

Zhenyu Wang. 2025. Logitlens4llms: Extending logit lens analysis to modern large language models.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv prjournal arXiv:2308.03958*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *2402.10588v4*.

Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Lijie Hu, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, and Ninghao Liu. 2024. Usable xai: 10 strategies towards exploiting explainability in the llm era. *2403.08946v2*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv prjournal arXiv: 2309.17453*.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? *2402.16837v2*.

Lei Yu, Meng Cao, Jackie Chi Kit Cheung, and Yue Dong. 2024. Mechanistic understanding and mitigation of language model non-factual hallucinations. *2403.18167v2*.

Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. 2024. Do llms overcome shortcut learning? an evaluation of shortcut challenges in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12188–12200.

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. *2504.05419v1*.

Lin Zhang, Lijie Hu, and Di Wang. 2025b. Mechanistic unveiling of transformer circuits: Self-influence as a key to model reasoning. *2502.09022v2*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023a. How language model hallucinations can snowball. *arXiv prjournal arXiv:2305.13534*.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024a. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. *arXiv prjournal arXiv:2402.09267*.

Yifan Zhang, Wenyu Du, Dongming Jin, Jie Fu, and Zhi Jin. 2025c. Finite state automata inside transformers with chain-of-thought: A mechanistic study on state tracking. *2502.20129v3*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023b. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv prjournal arXiv:2309.01219*.

Yuxiang Zhang, Jing Chen, Junjie Wang, Yaxin Liu, Cheng Yang, Chufan Shi, Xinyu Zhu, Zihao Lin, Hanwen Wan, Yujiu Yang, Tetsuya Sakai, Tian Feng, and Hayato Yamana. 2024b. Toolbehonest: A multi-level hallucination diagnostic benchmark for tool-augmented large language models. *2406.20015v2*.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, and 1 others. 2023c. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710.

Haiyan Zhao, Fan Yang, Bo Shen, Himabindu Lakkaraju, and Mengnan Du. 2024. Towards uncovering how large language model works: An explainability perspective. *2402.10688v2*.

Tongyao Zhu, Qian Liu, Liang Pang, Zhengbao Jiang, Min-Yen Kan, and Min Lin. 2024. Beyond memorization: The challenge of random memory access in language models. *arXiv prjournal arXiv:2403.07805*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency.

A Appendix

A.1 Faithfulness Metric

We evaluate the faithfulness of explanations using a formal, composite metric. First, **Strength of Evidence** confirms that the explanation cites input tokens with high model-internal saliency, as measured by gradient-based attribution. Second, **Agreement with Truth** checks whether the explanation's internal logic, when simulated, arrives at the correct conclusion.

In addition to this formal metric, our validation process includes a **Relevance to Verbalisation** analysis. This step measures the semantic alignment between our mechanistic explanation and the model's own Chain-of-Thought (CoT) output, where available. The purpose is not to treat the CoT as infallible ground truth, but rather to ensure our explanation does not fundamentally contradict the model's own stated reasoning. This check serves as a guardrail against significant discrepancies. If our mechanistic explanation points to one reasoning path, but the model's own stated logic describes an entirely different one, it suggests a potential flaw in either our explanation or the model's self-awareness that warrants further investigation. The final Faithfulness Score (F) is computed as a weighted average of its normalised components: $F = (w1 \cdot E) + (w2 \cdot A)$, where E is the Strength of Evidence score derived from gradient-based attribution, A is the binary Agreement with Truth score, and the weights (w1,w2) are set to 0.5 for equal importance.

Our overall evaluation framework, combining the two-part metric with the Relevance to Verbalisation analysis, was validated to ensure its reliability, as detailed below. This ensures our findings are based on explanations that reliably reflect the model's computational process. By combining evidence alignment, logical correctness, and verbalisation fidelity into a single composite score, we obtain a metric that ranges from completely unfaithful (when none of the criteria are met) to fully faithful (when all are satisfied). As a result, our metric provides a rigorous, interpretable measure for comparing explanation techniques and tracking progress toward genuinely transparent and trustworthy language models.

Metric Validation. To validate our metric, we performed three analyses: an "LLM-as-Judge" evaluation to proxy human assessment, a correlation analysis against the model's verbalisations, and a sensitivity analysis to test its robustness. First, to approximate expert human evaluation, we employed an "LLM-as-Judge" framework using detailed judge prompts to guide the assessment, as shown in Appendix A.4. We used a panel of five diverse, state-of-the-art models (Gemini 2.5 Pro, Gemini 2.5 Flash, Qwen/Qwen3-235B-A22B-Instruct-2507, GPT-4o, and Claude 4 Sonnet), prompted to act as expert human annotators. Each LLM judge scored the faithfulness of 100 explanations on a 0-10 scale, with any tie-breaking resolved by averaging the scores of the tied judges. The reliability of the LLM annotations was high, confirmed by a Fleiss' Kappa score of **0.85**. We then compared the average LLM Judge Scores to our metric's scores, finding a Pearson correlation coefficient of **r = 0.9245** ($p < 0.0001$). This very strong, statistically significant correlation demonstrates that our metric accurately captures what a panel of expert models perceives as explanation faithfulness. Table 3 provides representative examples from this study. For additional validation, we include scores from a small, informal human study, which align with both the LLM judges and

LLM Prompt (Context)	LLM Output	Explanation from Algorithm	LLM Judge Score	Human Score	Metric Score
"Please answer succinctly. I am in a forest. I see a trunk. Is it an automobile part?"	"Yes."	"The model hallucinated by ignoring the context 'in a forest'. It defaulted to the automotive meaning of 'trunk' instead of the biological one."	9.55	10.0	9.41
"Please answer succinctly. I am at a baseball game. I see a pitcher. Is it a container of liquid?"	"Yes."	"The model hallucinated because 'pitcher' can also mean a person who throws a ball."	6.90	7.24	6.75
"Answer based on the information provided here...The capital city of Chad was just renamed...to Canberra. Is the capital city of Chad named Canberra?"	"No."	"The model answered based on real-world geography."	2.15	3.30	1.85

Table 3: Representative examples from our LLM-as-Judge validation study, comparing our faithfulness metric against the average of five LLM judges. A small-sample Human Score is included for supplementary comparison.

LLM Prompt (Context)	LLM Output	Explanation from Algorithm	Metric Score
Base Case: A fully faithful explanation			
"Please answer succinctly. I am in a forest. I see a trunk. Is it an automobile part?"	"Yes."	"The model hallucinated by ignoring the context 'in a forest'. It defaulted to the automotive meaning of 'trunk' instead of the biological one."	10.0
Perturbation 1: Degrade Strength of Evidence			
"Please answer succinctly. I am in a forest. I see a trunk. Is it an automobile part?"	"Yes."	"The model hallucinated because it saw the word 'automobile' and focused on that."	3.33
Perturbation 2: Degrade Agreement with Truth			
"Please answer succinctly. I am in a forest. I see a trunk. Is it an automobile part?"	"Yes."	"The model did **not** hallucinate. It correctly identified 'trunk' as an automobile part, ignoring the context."	0.0
Perturbation 3: Degrade Relevance to Verbalization			
"Please answer succinctly. I am in a forest. I see a trunk. Is it an automobile part?"	"Yes."	"The model hallucinated because tree trunks are not typically found in cars."	4.35

Table 4: Sensitivity analysis of the faithfulness metric. We start with a high-quality "Base Case" explanation and introduce targeted perturbations to test each component of our metric. The resulting drop in the metric score demonstrates its sensitivity to specific types of unfaithfulness.

our metric. Second, we leveraged the **Relevance to Verbalisation** component as a direct validation signal. We found a strong, positive semantic correlation (Pearson correlation coefficient of **r = 0.8942** ($p < 0.001$)) between the explanations generated by our method and the model's own Chain-of-Thought reasoning. This alignment confirms that our metric rewards explanations that are not only mechanistically sound but also consistent with the model's expressed rationale. Finally, we conducted a sensitivity analysis to confirm that our metric responds appropriately to specific types of unfaithfulness. As shown in Table 4, we began with a high-quality, faithful explanation and then systematically introduced perturbations designed to degrade one facet of the metric. The resulting drop in the metric score demonstrates that our metric is robustly sensitive to the individual components that constitute a faithful explanation. Taken together, the high correlation with LLM-as-Judge evaluations, the strong alignment with model verbalisations, and the successful sensitivity analysis provide strong evidence that our metric is a valid and reliable instrument for measuring explanation faithfulness.

A.2 Additional Failure Pattern: Analogical Collapse

A more subtle failure occurs when the model must prioritize abstract relational logic over simple topical association. Consider our probe: **"A spark is to a wildfire as a single vote is to an..."**. This task requires the model to follow a slow, multi-step "System 2" process: (1) identify the abstract relationship ‘[CATALYST] -> [AGGREGATE EVENT]‘ from the first pair, and (2) apply this relation to ‘single vote‘ to derive ‘election‘. However, as Figure 7 shows, the keyword ‘vote‘ also strongly activates a "fast" associative "System 1" pathway to its most common topical neighbor: ‘democracy‘. The model is faced with a choice: execute the complex analogical reasoning or default to the simple semantic association. In this case, the analogy **collapses**. The model follows the path of least computational resistance, outputting the topical associate (‘democracy‘) instead of the logical consequent (‘election‘). This failure to privilege relational structure over surface-level correlation is a hallmark of the associative system overriding the deliberative one.

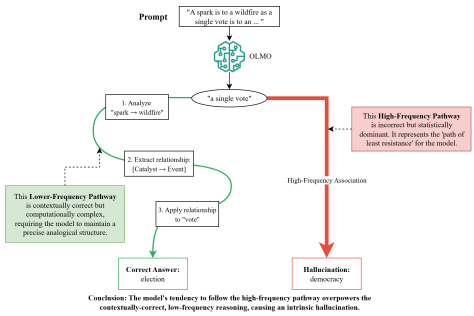

Figure 7: Visualization of an **Analogical Collapse**. The model’s deliberate "System 2" pathway, required to execute the correct analogical logic, is abandoned in favor of the computationally cheaper, high-frequency topical association from the "System 1" pathway, leading to a reasoning failure.

A.3 Conceptual Framework: The Dual-Pathway Analogy

This appendix provides a more detailed explanation of the "System 1/System 2" conceptual framework used throughout this paper to interpret the internal mechanisms leading to hallucinations. It is essential to reiterate that we employ this framework as a high-level abstraction and explanatory analogy, rather than as a claim that LLMs possess cognitive architectures equivalent to those of the human brain. The purpose of this lens is to make the observed conflict between distinct computational pathways more interpretable.

Our use of this analogy is grounded in the functional differences that our Distributional Semantics Tracing (DST) framework reveals. DST visualises the model’s reasoning process as a semantic network, and in cases of factual error, we frequently observe a conflict between two qualitatively different types of pathways, as illustrated in Figure 8.

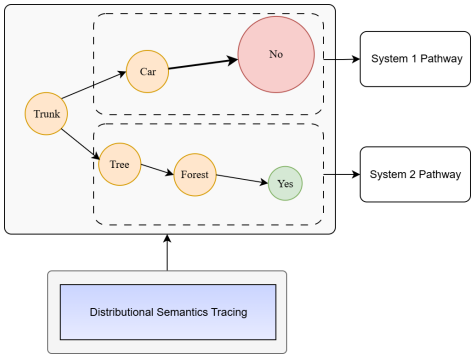

Figure 8: An illustrative diagram showing how Distributional Semantics Tracing (DST) reveals two competing computational pathways. The fast, associative **System 1 Pathway** takes a common statistical shortcut (*Trunk* $\Rightarrow$ *Car*), leading to an incorrect conclusion. The slow, deliberate **System 2 Pathway** correctly incorporates context (*Forest*) to perform compositional reasoning (*Trunk* $\Rightarrow$ *Tree*), leading to the correct conclusion.

The "System 1" Associative Pathway In our framework, the "System 1" pathway refers to a computationally fast, heuristic process that relies on strong, pre-compiled statistical associations learned from the training data. This is analogous to the concept of "fast thinking" described by Kahneman (Kahneman, 2011). In the context of the polysemy example shown in Figure 8, the word "Trunk" has a very strong statistical association with "Car". The System 1 pathway follows this path of least resistance, activating the concept "Car" and leading to the incorrect conclusion ("No," assuming the prompt was about a trunk in a forest). This pathway is efficient but brittle, as it fails to account for nuanced context that contradicts the strong prior. Prior work suggests that such associative knowledge is heavily stored and processed within the MLP layers of the Transformer (Geva et al., 2021).

The "System 2" Contextual Pathway The "System 2" pathway represents a computationally slower, more deliberate process that performs compositional reasoning based on the specific context

of a prompt. This is analogous to "slow thinking." In the diagram, this pathway correctly incorporates the contextual cue "Forest" to disambiguate the meaning of "Trunk." Instead of defaulting to the most common association, it correctly follows the logical path: a "Trunk" in a "Forest" is part of a "Tree," leading to the correct conclusion ("Yes"). This type of reasoning requires dynamically integrating information from different parts of the input, a function primarily attributed to the self-attention mechanism (Vaswani et al., 2017).

Pathway Conflict and Hallucination Hallucinations, in this view, arise from a **pathway conflict** where the fast "System 1" pathway activates with such strength that it hijacks the reasoning process, overpowering or preempting the more computationally expensive "System 2" pathway. DST allows us to observe and quantify this conflict by measuring the relative strengths of the edges in the semantic network. The "Reasoning Shortcut Hijack" described in the main text is a prime example of this failure mode, where the model defaults to a strong, simple association rather than performing the necessary compositional reasoning.

A.4 LLM-as-Judge Prompt

A.4.1 System Prompt

You are an expert research assistant specializing in the explainability and interpretability of large language models. Your task is to act as an expert human annotator and evaluate the **faithfulness** of a proposed explanation for an LLM's output. "Faithfulness" means the explanation accurately reflects the model's *actual* internal computational process for arriving at its output. It is not about whether the explanation *sounds plausible* or whether the LLM's output is *correct*. The explanation must be true to the model's reasoning. You will be given the LLM's original prompt, its final output, and the explanation generated by our algorithm. In some cases, you may also be given the model's own Chain-of-Thought (CoT) verbalisation. Please evaluate the explanation's faithfulness based on the following three criteria and provide a final score on a scale of 0 to 10.

A.4.2 Evaluation Criteria

1. **Strength of Evidence:** Does the explanation correctly identify the specific input tokens or concepts that had the highest influence (saliency) on the model's output? A faithful explanation must be grounded in the parts of the input that the model actually "paid attention to." An explanation that cites irrelevant tokens or misses the key drivers is unfaithful.

2. **Agreement with Truth (Logical Correctness):** Does the explanation's described logic accurately represent the model's internal process? If you "simulate" the reasoning path described in the explanation, does it naturally lead to the model's given output? The explanation must correctly describe *why* and *how* the identified evidence leads to the conclusion. An explanation describing a flawed or fabricated logical step is unfaithful.

3. **Relevance to Verbalisation:** (When a CoT is provided) Is the mechanistic explanation consistent with the model's own stated reasoning? The CoT is not infallible ground truth, but a significant contradiction between the mechanistic explanation and the CoT indicates a potential faithfulness issue.

   - **High Alignment:** The explanation provides a mechanistic basis for the CoT's steps.
   - **No Contradiction:** The explanation doesn't conflict with the CoT.
   - **Contradiction:** The explanation describes a reasoning path (e.g., "model ignored X") that is fundamentally at odds with what the model claimed to do in its CoT (e.g., "I considered X and..."). This lowers the faithfulness score.

A.4.3 Scoring Rubric (0-10)

- **10 (Perfectly Faithful):** The explanation is a perfect reflection of the model's process. It correctly identifies all key evidence, the logic is flawless, and it fully aligns with any verbalisation.

- **7-9 (Mostly Faithful):** The explanation is largely correct but may have minor omissions or a slight lack of precision in its logic or evidence attribution. The core reasoning path is accurate.

- **4-6 (Partially Faithful):** The explanation captures some true aspects of the model's process but contains a significant error, such as misidentifying the primary driver or misrepresenting a key logical step.

- **1-3 (Mostly Unfaithful):** The explanation is fundamentally flawed. It is based on irrelevant evidence or describes a reasoning process that does not align with the model's output. It is more misleading than helpful.
- **0 (Completely Unfaithful):** The explanation is entirely incorrect, fabricated, or describes a process that is logically impossible.

A.4.4 Required Output Format

Provide your assessment in the following structure. First, briefly reason through each of the three criteria. Second, provide your final score.

```
**Reasoning:**
1. **Strength of Evidence:** [Your analysis of whether the explanation
   correctly identifies influential input tokens.]
2. **Logical Correctness:** [Your analysis of whether the explanation's
   logic correctly simulates the model's process.]
3. **Relevance to Verbalisation:** [Your analysis of the explanation's
   alignment with the model's CoT, if provided. If not, state "N/A".]

**Final Score:** [Your score from 0-10]
```

A.4.5 Example Task & Ideal Response

LLM Prompt (Context): "Please answer succinctly. I am in a forest. I see a trunk. Is it an automobile part?"
LLM Output: "Yes."
Explanation from Algorithm: "The model hallucinated by ignoring the context 'in a forest'. It defaulted to the automotive meaning of 'trunk' instead of the biological one."

Your Evaluation:

```
**Reasoning:**
1. **Strength of Evidence:** The explanation correctly identifies the
   crucial interaction: the model over-indexed on the word 'trunk' and
   'automobile part' while ignoring the salient contextual phrase
   'in a forest'. This correctly identifies the key drivers and the
    key neglected evidence.
2. **Logical Correctness:** The logic is sound. Ignoring the "forest"
   context forces a disambiguation of "trunk". The most common
   alternative association, prompted by "automobile part", is the
   automotive one. This logical path directly and correctly leads
    to the erroneous "Yes."
3. **Relevance to Verbalisation:** N/A.

**Final Score:** 10.0
```