Language Without Propositions: Why Large Language Models Hallucinate

Jakub Mácha *

Posted Date: 20 January 2026

doi: 10.20944/preprints202601.1447.v1

Keywords: large language models; artificial intelligence; proposition; fact; truth; hallucination; logical atomism; radical interpretation

 doi:10.20944/preprints202601.1447.v1

Article

Language Without Propositions: Why Large Language Models Hallucinate

Jakub Mácha

Masaryk University, Czech Republic; macha@mail.muni.cz

Abstract

This paper defends the thesis that LLM hallucinations are best explained as a *truth representation problem*: Current models lack an internal representation of propositions as truth-bearers, so truth and falsity cannot constrain generation in the way factual discourse requires. It begins by surveying leading explanations—computational limits on self-verification, deficiencies in training data as truth sources, and architectural factors—and argues that they converge on the same underlying representational deficit. Next, it reconstructs the philosophical background of current LLM design, showing how optimization for fluent continuation aligns with coherence-style evaluation and with a broadly structuralist, relational semantics, before turning to David Chalmers's recent attempt to secure propositional interpretability by drawing on Davidson/Lewis-style radical interpretation and by locating propositional content in "middle-layer" structures; it argues that this approach downplays the ubiquity of hallucination and inherits instability from post-training edits. Finally, the paper offers a positive proposal: Atomic propositions should be represented in the basic vector layer, reviving a logical-atomist program as a principled route to reducing hallucination.

Keywords: large language models; artificial intelligence; proposition; fact; truth; hallucination; logical atomism; radical interpretation

0. Introduction

Steven Pinker, in a recent (June 2025) conversation with Richard Dawkins, made the following remarks on the shortcomings of LLMs:

> I think the shortcomings of neural network models are still there, and they're called hallucinations or confabulations or blends, and it's precisely because their intelligence comes from generalizing based on the similarity to things in the training set, and blending things of the general kind that tend to go together, even when we now know they never did go together, and hence you get the hallucinations. Several years later, still a problem, and as a semi-regular user of generative AI, I'm still surprised at the hallucinations that come about, and they come about for a systematic reason.
>
> Namely, there's nothing in there corresponding to a proposition, to the capital of France is Paris, or so-and-so did such-and-such at such-and-such a time. There are blends of many things that tend to co-occur in the training set, resulting in output that is always plausible, but not always factual. (Dawkins and Pinker 2025, 1:03:13–1:04:24)

The aim of this paper is to provide a rigorous elaboration of the idea expressed by Pinker: *The systematic reason why LLMs hallucinate is that they do not have any account of the propositions that could convey true factual knowledge.* I will clarify this initial formulation once the necessary context has been provided.

What is hallucination with respect to LLMs? A recent survey study by Manuel Cossio provides us with the following useful definition: Hallucination "signifies the creation of nonfactual information to respond to a user's query, frequently without any explicit indication of its fabricated

 doi:10.20944/preprints202601.1447.v1

nature" (Cossio 2025, 4). For our present purposes, it is important that hallucination is concerned with factual knowledge or information. "Factual" in this context means "true." Hallucinating LLMs thus produce "false or fabricated information" (ibid., 5). Among several categorizations of hallucination set out by Cossio, I shall focus on the so-called "factuality hallucination," which can be characterized as contradicting "real-world knowledge."

The main idea I want to defend in this article is that LLMs tend to hallucinate because they do not have any inherent account of truth (as correspondence to facts). LLMs represent meaning on the level of tokens, whereas truth (or falsity) pertains to sentences or propositions. I will refer to this issue as LLMs' *truth representation problem*.

The truth representation problem is preceded by another related yet distinct issue: LLMs are not provided with true propositions that are reliably marked as such within their input data. LLMs receive inputs of two types: training data and user prompts. The training data are not reliable sources of true facts. Although true information frequently occurs (and perhaps prevails) among training data, training data cannot be considered a reliable arbiter of truth. (Even if human fine-tuning can eliminate the most blatant cases of falsity.) A similar consideration applies to user input: A user may be wrong or even try to mislead the LLM. Moreover, user input often includes results of an internet search, which are not under the user's direct control. In addition to this, there is the so-called "system prompt": a sometimes public, but often hidden addendum to the user's prompt imposed by the LLM's operator (usually a tech company, such as OpenAI). All in all, none of these kinds of input data are a reliable source of truth. Even if LLMs had an internal representation of factual veracity, they would not have any reliable way to verify whether some purported fact is indeed true. However, I do not want to claim that hallucinations only occur because of false input data. Falsity itself is not the main problem. Rather, it is LLMs' inability to distinguish between truth and falsity. I will refer to this issue as *the truth source problem*.

Both problems are truth-related. However, they are mutually independent issues. Even if LLMs are given true and only true propositions as their training set, they may still struggle to accurately represent such propositions. And, conversely, even if LLMs could represent true propositions, this ability would be good for nothing if they were not provided with any true propositions in the first place. To put it differently: If we could adequately address the truth representation problem, LLMs would be *models of language* capable of expressing all possible facts. If we were to additionally solve the truth source problem, LLMs would be adequate *models of the world* taken as the totality of all actually true propositions.

LLMs' hallucinations usually do not directly contradict information that occurs in their training dataset. The typical issue is that they fabricate and assert with apparent certitude a claim that does not occur in their training set. This kind of hallucination, which I will refer to as *confabulation*, is independent of the truth source problem, but rather occurs due to the truth representation problem. Or so I want to argue.

Is "Hallucination" a Fitting Metaphor?

Before I dig into the argument proper, let us briefly consider the rationale of the term "hallucination." Simplifying somewhat, we can say that the term originates in folk psychology. It is a subjective phenomenon of experiencing involuntary impressions in the absence of any actual perception—e.g., hearing voices or seeing things that are not really there (that is, are not actually affecting the subject's senses). Hallucination is a cognitive defect caused by mental illness, certain drugs, poisoning, or lack of sleep. The term primarily has these negative connotations. When I use the term in the context of LLMs, it serves as a metaphor to loosely express the same basic idea, though not necessarily all the connotations that go with it. The core of the metaphor appears to be this: Just as psychological hallucinations refer to content without corresponding sensual stimulations. LLMs' hallucinations are factual statements without corresponding facts. In short, psychological hallucinations are false impressions, whereas LLMs' hallucinations are false statements.

 doi:10.20944/preprints202601.1447.v1

However, the hallucination metaphor can be misleading, because it also suggests that LLMs can perceive and possess internal phenomenological states. This is to say, the metaphor unnecessarily anthropomorphizes AI systems (cf. Weatherby's (2025) discussion of "remainder humanism"). It encourages users to attribute agency, belief, or conscious error to the systems. And yet there are even more serious issues with the metaphor, as it conflates propositions with percepts, which detracts attention from the fact that the core of the problem lies in LLMs' representation of propositions. We are led to think that the model *knows* the truth and is merely misperceiving it. Finally, the metaphor also suggests that the issue is a marginal bug. Real hallucinations are relatively rare occurrences and not a fundamental flaw in human psychology. In contrast, as I shall argue, the hallucinations seen in LLMs stem from a basic architectural flaw in the models' design.

There seems to be a battery of more suitable terms that avoid some of these misleading implications. Pinker already mentioned "confabulation." Another possibility is "fabrication." Other, perhaps less cogent terms include "ungrounded generation," "propositional error," "referential failure," and "model divergence." The only reason I will continue to use "hallucination" in this paper is because it has become the customary term and despite the issues described here nevertheless refers to the problem I wish to address.

As already made clear, the term "hallucination" has rather negative connotations. However, even within human psychology, we can refer to similar phenomena using positive terms, such as "imagination" or "fantasy." It would be equally misleading to say that LLMs possess imagination. The phenomena that humans experience have both positive and negative aspects. This leads us to the question of whether LLMs' hallucinations have positive aspects as well. We must realize that there are discourses and contexts where truth or factual accuracy is not the ultimate goal. In human communication, factual statements are often embedded in pragmatic contexts that modify or even cancel their reference to corresponding facts. In the context of art and fiction, for instance, fluidity can be more valuable than factual accuracy. LLMs' tendency to hallucinate can be instrumental when they are employed in such contexts or for such tasks.[1] Hence, the present argument should not be taken to imply that LLMs' hallucinations are entirely bad in all cases and must always be eliminated.

Hallucination in LLMs resembles *bullshit* in Harry Frankfurt's sense, because both arise from a disregard for truth as such. Frankfurt (2005) distinguishes the liar, who still orients themselves toward truth, even if only to conceal it, from the bullshitter, who simply does not care whether what they say is true or false. The bullshitter's utterances aim not at representing reality but at producing a desired impression. LLM hallucination operates under a structurally analogous indifference: The model's objective is not to assert truth but to generate a coherent, contextually appropriate continuation of text. When a model hallucinates, it is not *mistaken* about the world, but altogether *indifferent* to it. Its training optimizes for plausibility and fluency, not for correspondence to facts. The resulting utterance may, by coincidence, be true, but its truth or falsity is epistemically irrelevant to the generative process.

There is, however, an essential difference between a bullshitter and a hallucinating machine. Bullshitters could, if they wished, speak truthfully—their indifference is an *attitude* toward a norm they understand. In other words, the truth representation problem is not an issue for the bullshitter. An LLM, by contrast, has no concept of truth and therefore no capacity to disregard it. Its indifference is not chosen but built in: a consequence of optimization for prediction rather than assertion. What in Frankfurt's human case marks an ethical and epistemic vice—inattention to a known standard—becomes in the machine case a structural feature of design, namely, the truth representation problem. Thus, while LLM hallucination and bullshit share the same outward form—language unconstrained by truth—their underlying logics diverge. Human bullshit is a performative stance toward truth; machine hallucination is a semantic void in which the very notion of truth never enters the process.

---

[1] Weatherby (2025) addresses these positive functions under the headings of "general poetics" and "poetic ideology." Cf. also Lee and Mácha (2024) for a discussion of the ramifications of the hallucination metaphor for LLMs' ability to create works of art.

 doi:10.20944/preprints202601.1447.v1

The paper is organized as follows. Section I begins with a review of the customary explanations of hallucinations in the recent literature, culminating with my preferred explanation: namely, insufficient representation of propositions. Sections II and III flesh out the philosophical ideas underlying the design of current LLMs: a preference for the coherence theory of truth (sect. II) and a structuralist account of language that prioritizes purely relational semantics (sect. III). In section IV, I turn to a recent proposal put forward by David Chalmers, which aims to represent propositions in the so-called "middle layer." I argue that Chalmers's attempt to base his proposal on Donald Davidson's radical interpretation is insufficient precisely because it downplays the problem of hallucinations. Section V presents my positive proposal to represent propositions in the basic layer of vectors, together with a discussion of the underlying account of language: logical atomism.

I. Previous Explanations of LLMs' Hallucinations

The issue of hallucination has attracted significant scientific and academic scrutiny. Various explanations for the underlying causes of hallucination have been proposed, which can be categorized into three main groups: computational, training-data-related, and model-architecture-related.

(1) The first and most rigorous explanation is rooted in the theory of computation (Xu et al. 2024; cf. also Cossio 2025, sect. 2.2). It can be proven, using the technique of diagonalization, that an LLM will always hallucinate on infinitely many inputs. This argument construes a "truth ground function," which for each input string gives the only correct output. Using the diagonalization technique, we can construct a truth ground function in such a way that each LLM will hallucinate with respect to it. The conclusion is striking: No LLM can self-referentially assess the veracity of its outputs and thus prevent itself from hallucinating.

The mathematical proof is undoubtedly correct within its idealized formal framework. If there were a real-world counterpart to the truth ground function, then LLMs would inevitably hallucinate regarding it. My concern, however, is that current LLMs are unable to compute a truth ground function at all. If the primary cause of hallucination were purely computational, I would question why LLMs still hallucinate in tasks that are unrelated to self-reference, which is a factor in the diagonalization technique. In other words, while the theory of computation can explain some instances of hallucination, it does not account for others.

(2) The training datasets are the primary sources of all the content LLMs have at their disposal. It is natural to suppose that any flaws and biases in the training data could reappear in the LLM outputs. During the training stage, when a model is being composed, it is trained on a vast amount of textual data obtained from the internet and other sources.[2] At least during this phase, the source data are not verified for their factual accuracy, relevance, or any potential biases. In line with recent literature (Joshi 2025; cf. Cossio 2025, 16), we can identify three sorts of issues pertaining to the training data: (i) flawed information content (including false, incomplete, noisy, or contextually irrelevant data), (ii) biased data, which includes all sorts of biases: racial, gender, religious, political, etc., (iii) outdated training data: information that was once considered true but has since been disproven. The *truth source problem* described above covers all three issues.

Since all that matters, and all that is extracted from the source dataset, is statistical patterns between words/tokens, statistical harvesting can actually amplify these issues. Let us focus on a few examples. Temporal knowledge analysis by Jang et al. (2022) shows that pre-2020 large corpora

[2] Large-scale foundation models are typically trained on mixtures of uncurated web-scale corpora—such as Common Crawl, C4, OSCAR, and the Pile—together with curated factual datasets like Wikipedia, government publications, and filtered scientific and educational resources. Uncurated corpora are *massive web scrapes* containing heterogeneous, duplicate, noisy, or incorrect material; their scale is crucial for model performance, but they are not verified for accuracy. By contrast, curated corpora undergo explicit editorial or algorithmic filtering for quality and topical relevance.

 doi:10.20944/preprints202601.1447.v1

usually contained far more outdated scientific claims, such as "Pluto is the ninth planet" statements rather than "Pluto is a dwarf planet." Another study using large-scale textual corpora indicates that the common myth "Humans use 10% of their brain" occurs at a substantially higher textual frequency than formulations expressing the true proposition that they use their full brain capacity (Michel et al. 2011). Another plausible real-world example of a false claim outweighing its true negation in public discourse is the assertion that vaccines cause autism, which has repeatedly been shown to dominate attention, circulation, and engagement relative to scientific information debunking that claim (Broniatowski et al. 2018; Jolley and Douglas 2014).

(3) One of the tasks in the *post-training* phase is to correct the model's factual inaccuracies. This endeavor presupposes that hallucination emerges due to the *truth source problem*. As this is a complex issue related to model architecture, we will revisit it in more detail later. As it stands, this data-related argument suggests that LLMs are prone to hallucinating with regard to issues originating in the training dataset. This is, without doubt, one of the causes of hallucination. One way to mitigate it would be to use more reliable training data or to perform some sort of manual curation (before the training or in the post-training stage). However, as already indicated above, actual cases of hallucination usually involve propositions not present in the training data. One could object that it is difficult to back this claim empirically. But for the moment, I can say that I am interested in this kind of hallucination without claiming that it occurs more often than biases inherited from the training data (we will revisit this point in due course). The cause of this kind of hallucination must lie in the model architecture, to which we shall now turn.

(4) Various architecture-related causes of hallucination have been proposed in the literature (for an overview see Cossio 2025, sect. 5.2). These can be summarized in the statement that LLMs are not designed to produce truthful output. Their primary goal is not factual accuracy, but rather to predict the most probable token. However, these two tasks often go hand in hand. Following statistical patterns can, and in many cases does, lead to factual accuracy. To give a simple example: Take the input "The capital of France is …" The most probable next token is "Paris," and that output is also factually accurate. This example, however, is oversimplified, because it presents a case of a sentence occurring verbatim in the training dataset. As already indicated, I aim to focus on cases where outputs extend beyond replicating the source data, since there is nothing particularly "intelligent" in such replication. To identify the underlying causes of these types of hallucinations, we must examine LLMs' internal design in detail.

One plausible culprit could be setting the *temperature* parameter too high (Huang et al. 2025). Obviously, if the model does not generate the most probable next token sometimes (depending on the temperature), it is prone to returning low-probability outputs, which are more likely to be false (this is just the argument from the previous paragraph reversed). However, setting the temperature parameter too high is not an architecture flaw, but rather the user's choice (more precisely, it is a parameter of the user interface through which the model is accessed). An obvious remedy would be to set the temperature parameter to zero. The model would become completely deterministic as only the single most likely token would ever be chosen. However, this solution does not work in practice. The model would not be able to "explore" continuations that are slightly less probable but possibly more correct or contextually appropriate. Setting temperature to zero makes the model rigid, brittle, and overconfident, sacrificing robustness and adaptability for determinism. A small, nonzero temperature (e.g., 0.2–0.5) usually yields better factual and stylistic performance because it allows the model to "hedge" against its own uncertainty and choose more contextually appropriate continuations. This suggests that the model's inherent randomness, parametrized by temperature, is not the primary cause of hallucination.

Another suggested cause of hallucination is flaws in the *softmax* algorithm, which converts a list of raw scores into a probability distribution (Chang and McCallum 2022). Because softmax represents all possible word probabilities as a low-rank matrix, it cannot fully capture the rich, complex distribution of natural language, which leads to restricted expressiveness. This fundamental limitation has been referred to as the "softmax bottleneck." Since the softmax layer amplifies some

 doi:10.20944/preprints202601.1447.v1

probabilities while curtailing others, it can only amplify already-existing instances of hallucination. This leaves it unexplained how these instances originally occurred.

(5) It has been suggested that actual hallucinations about facts not present in the training data can arise due to LLMs' limited reasoning capacities. This is a variation on the issue already addressed above. LLMs' primary goal is token prediction, not logical inference. LLMs do not possess any inherent logic; they are not grounded in any fundamental axioms or laws of thought. If LLMs are capable of logical reasoning, it is a by-product of their primary operation, namely, predicting the next token. We could adjust the same argument as we did with "The capital of France is …" to, say, the *modus ponens* "If *P* implies *Q*. *P* is true. Therefore, …" In that case, the most probable next token is likely to be *Q*. However, this instance of *modus ponens* can also occur in the training data. LLMs will more likely hallucinate on complex chains of reasoning that do not occur there. This is, without doubt, a crucial issue.

In my view, this is a consequence of a more fundamental issue: the lack of knowledge representation in LLMs. Logical reasoning involves reasoning about knowledge represented by propositions (as *P* and *Q* stand for propositions in my example). In our simplified *modus ponens* case, the LLM is asked to manipulate symbols, including propositional variables *P* and *Q*. In this case, the LLM does not need to know what *P* and *Q* stand for. In real-life scenarios, LLMs can be utilized to process and work with natural-language sentences. For example, "If it rains, the path will be wet. It is raining. Therefore, …" In order to be able to process this inference, the LLM must *understand* that "It is raining" is a representation of a true proposition. We can say that logical reasoning preserves truth because, in a valid argument, if the premises are true, then the conclusion must also be true by virtue of the argument's *form*, not its content. This means that the LLM must be able to ascertain the truth of the antecedent (here: "It rains"). Thus, the capacity to accurately represent (true or false) propositions is essential for logical reasoning.[3]

(6) The next cause, though not the primary one, of hallucination discussed in scholarly literature is extensive *optimization and model fine-tuning*. This might seem surprising, as a primary goal of fine-tuning is to enhance a model's factual accuracy and reduce its hallucination rate. However, there has been growing empirical evidence that when fine-tuning is optimized too aggressively in one domain, it can lead to an increased hallucination rate in other adjacent domains. More specifically, when optimized for task-specific metrics (e.g., BLEU, ROUGE, accuracy, reward models), the model can become overconfident in spurious correlations and generate fluent but false content. This is consistent with evidence that instruction-tuning and reinforcement learning from human feedback (RLHF),

[3] Current LLMs (like GPT-5 and Grok 3) utilize what is known as the *chain-of-thought* method: They display the chain of their logical reasoning, from the input (user prompt, together with additional web searches) through intermediate steps to the output. This feature was introduced to make reasoning explicit and reveal how the model moves from premise to conclusion. The method assumes that natural-language sentences can stand for propositions whose truth can be determined, and that displaying intermediate steps allows the user to inspect the structure of inference. The model, so to speak, thinks aloud. The user is given the impression that the model arrived at its output via a rigorous logical reasoning process. This process structurally resembles human thinking. In theory, this bridges linguistic expression and logical reasoning. In practice, it does not. For as we know, this is not how LLMs actually reason. The chain-of-thought display is an explanatory fiction produced for the user. This fiction is useful to the extent that the user can review it and correct the model if it does not proceed in the desired direction. But the method also has serious issues and limitations, ranging from restricted generality (Chalmers 2025) to making outright false claims, that is, producing hallucinations (Turpin et al. 2023). The method as a whole can be viewed as a single, grand hallucination, as it gives the user a false impression of rigorous logical reasoning at the LLM's core. Moreover, the user is misled into thinking that the model can represent propositions, which, in my view, is the main flaw in its design.

 doi:10.20944/preprints202601.1447.v1

while improving helpfulness, can increase confident hallucinations by encouraging models to always produce an answer, even when uncertain.[4]

The recurring pattern is thus that, in the post-training (fine-tuning) phase, the model is adjusted to produce certain outputs (or to increase the probability of producing them) or not produce other outputs (typically those that are judged by human fine-tuners to be factually wrong or inappropriate). However, while this is achieved, the probability of other outputs, those not directly focused on by fine-tuning, is altered. Let me illustrate this with examples that have already been reported in the literature. De Cao et al. (2021) have demonstrated that post-training correcting the false factual statement "The capital of Namibia is Namibia" into the correct statement "The capital of Namibia is Windhoek" has led to a reduction in the probability of predicting "Moscow" when asked "What is the capital of Russia?" This shows that editing language models to correct specific factual errors (e.g., updating political office holders) increases accuracy on those targets but simultaneously lowers accuracy on related, unedited facts. Similarly, Meng et al. (2023) and Mitchell et al. (2022) have shown that even successful factual corrections can propagate distortions to semantically neighboring entities.

What this means is that fine-tuning that turns a falsehood into a truth can also induce new hallucinations through *representational drift*. Hence, fine-tuning does not just store a new key–value pair; it resculpts a whole region of the model's vector space. The reason for this is not philosophical, but primarily computational or mathematical: LLMs are not endowed with data structures for storing factual knowledge in the form of fact–truth value. In simpler terms, there is no data structure capable of accurately representing these propositions. This brings us back to Pinker's objection: "There's nothing in there corresponding to a proposition." Any attempt to embed factual information into an LLM can lead to distortion and subsequent hallucination, as there is no dedicated place for this information to be stored. Propositional knowledge is simply data in the wrong format; propositions are the wrong currency for LLMs (in their current design).

(7) This discussion of the issues with fine-tuning leads us to my main claim: *The primary cause of LLMs' hallucination is that they lack a direct account of propositional knowledge that can be evaluated for its truth or falsity.* Although a growing number of primarily empirical studies[5] have supported this

[4] Empirical evidence supports the view that extensive or poorly targeted fine-tuning can actually *increase* hallucination rates in LLMs. Gekhman et al. (2024) show in controlled closed-book QA experiments that since fine-tuning data contain a higher proportion of previously unknown or novel facts, a model's propensity to hallucinate relative to its original knowledge base rises almost linearly, with early stopping only partially mitigating the effect. Ghosal et al. (2024) similarly found that fine-tuning LLaMA-7B and Mistral on *low-popularity* factual data worsens performance by roughly 7–10 percent on factuality benchmarks such as PopQA and MMLU, demonstrating that narrow or low-coverage fine-tuning can degrade a model's general truthfulness. Lin et al. (2024) further observe that conventional alignment and reward-model objectives, optimized for fluency, helpfulness, and verbosity, tend to over-encourage long and confident answers, thereby amplifying plausible but unfounded statements; their proposed factuality-aware loss function improves this but confirms the baseline bias. For rare or low-frequency knowledge, approaches that rely on external information sources tend to be more robust than direct fine-tuning on sparse data, since narrowly targeted fine-tuning can overfit to limited examples and introduce additional factual errors. Finally, Zhang et al. (2024; 2025) have empirically established a *law of knowledge overshadowing*: Hallucination frequency grows with data imbalance, knowledge popularity, and model size, as dominant facts "overshadow" rarer ones in gradient updates. Collectively, these studies support the conclusion that while fine-tuning enhances local adaptation, over-optimization, data imbalance, and alignment biases can systematically *raise* hallucination rates outside the tuned domain.

[5] A number of recent studies advance or reinforce variants of the claim that the primary cause of hallucination lies in the absence—or at least the instability—of propositional, knowledge-bearing representations within LLMs.

 doi:10.20944/preprints202601.1447.v1

claim, my argument is conceptual, drawing on both the nature of LLMs' architecture and philosophical reflection on related epistemological and linguistic issues. The argument goes as follows: Hallucination is primarily a failure to produce truthful output. To avoid it, that is, to produce only truthful claims, LLMs must be able to represent truth in their internal structure. In other words, *hallucination is the truth representation problem*. Now, in order to represent truth, one has to be able to represent the truth-bearer, which is the proposition.

One might wonder why I insist on the notion of proposition so vehemently. There can be nonpropositional truth or even nonpropositional knowledge. Indeed, such concepts are increasingly being adopted. However, they are not relevant to characterizing LLMs' hallucinations. Let us take, for instance, Heidegger's notion of truth as un-concealment. On this account, hallucination would be concealment. However, it is not plausible to say, at least in the literal sense, that LLMs conceal something if they are aiming to generate the most probable next token. A similar consideration goes for nonpropositional knowledge-how, as proposed for example by Ryle (1949). Here, it is difficult to imagine how LLMs can be endowed with that kind of practical knowledge at all. Hallucination, on this account, would be a failure to act on such knowledge. However, LLMs do not act in the world (they have no practical abilities). They only generate tokens within their fixed symbolic system. Hence, we can, for present purposes, say that hallucination is a failure to produce veridical propositional content.[6]

At the same time, I do not want to claim that an adequate truth representation would solve the hallucination problem. Rather, it is one crucial step toward solving this problem. Truth representation

---

Chen et al. (2024) approach the problem from an *inner-representation* perspective, showing that hallucinated outputs correlate with diffuse and nondiscriminative activation patterns, implying that the model fails to encode distinct propositional states. Chekalina et al. (2024) demonstrate that supplementing LLMs with *knowledge graph embeddings* markedly reduces hallucinations, effectively treating the phenomenon as a representational deficiency remediable through the injection of structured propositional content. Similarly, Sansford et al. (2024) propose the GraphEval framework, which evaluates factuality at the level of propositional triples—implicitly assuming that hallucination is a breakdown in proposition-level encoding. Zhang et al. (2025) frame the problem as one of "knowledge overshadowing," where existing information is either incompletely represented or overwritten by spurious associations, again situating hallucination in defective internal knowledge representation. Finally, broader analyses of knowledge graphs and LLMs (Lavrinovics et al. 2024) converge on the view that language models hallucinate because they lack stable propositional anchoring to factual content—an insufficiency only partially offset when external symbolic structures are integrated. These works lend empirical and conceptual weight to the thesis that hallucination stems not merely from data bias or decoding artifacts, but from a model's failure to instantiate internally coherent, truth-evaluative propositions.

[6] In line with a recent article by Chalmers (2025; addressed in detail in section IV), I can admit that "pictorial or map-like representations" can also be truth-bearers, although they are not propositions proper. Chalmers treats such structures together with propositions under the heading "generalized propositional attitudes". However, LLMs are primarily textual machines, and their hallucinations are textual, which means, according to my argument, that LLMs' hallucinations are primarily related to propositions. Other promising current AI systems are visual language models (VLMs) whose hallucinations are primarily pictorial. Recent architectural designs, such as DeepSeek-OCR (Wei et al. 2025), combine textual and visual language models. In such systems, representing generalized propositions would be crucial to combat the hallucination problem. Taken from another perspective, there are accounts of propositions that are pictorial in essence—Wittgenstein's *Tractatus* being the most seminal one. What is crucial here is the ability to represent factual content that may or may not be accurate. Hence, I can keep insisting on the centrality of the proposition without diminishing the pictorial dimension of representation.

 doi:10.20944/preprints202601.1447.v1

is, so to speak, a necessary requirement (*conditio sine qua non*) of avoiding hallucinations. Moreover, most of the other suggested causes of hallucinations are related to the truth representation problem in one way or another. Solving the truth source problem would be for nothing if the model could not represent true knowledge originating from a reliable source—be it training data or a fine-tuning method. And as previously stated, if propositions are fundamental components of logical inferences, then the ability to engage in reliable logical reasoning requires the ability to represent propositions.

II. The Debate About Hallucination Echoes the Tension Between Coherence and Correspondence

If we view hallucination primarily as a failure to accurately represent and ground truth, we can link this issue to the traditional debate surrounding truth grounding. Hence, the current debate about hallucination and grounding in LLMs is, in philosophical terms, a direct spinoff of the long-standing tension between coherence and correspondence theories of truth. At its core, the question of whether an LLM's outputs can be considered "true" without reference to the external world rehashes the classical dispute between truth as internal consistency and truth as alignment with reality. The coherence view sees a model's linguistic competence—its ability to maintain syntactic, semantic, and pragmatic consistency within a self-contained web of textual relations—as sufficient for not only meaning, but also truth. This echoes the idealist claim that truth resides in the harmony of beliefs. The opposing, correspondence-based stance insists that such internal coherence must be *grounded* in contact with facts, perception, or causal interaction with the world, otherwise the model remains a closed formal system producing plausible but unanchored discourse. In this sense, contemporary discussions of symbol grounding, hallucination, and referential opacity in LLMs are not addressing new philosophical problems but rather are modern reformulations of the traditional question of whether coherence alone can yield truth or whether reality must have the final word.

While philosophers have long recognized that coherence is a necessary condition for truth—since self-contradictory propositions cannot all be true—the mainstream view today is that coherence alone is not sufficient. As Russell observed, "a whole system may be perfectly self-consistent, and yet entirely false. A novel, for example, may be self-consistent, but it is not therefore true" (Russell 1912, 132). This intuition continues to guide most contemporary theories of truth and knowledge. Even thinkers sympathetic to epistemic holism, such as Donald Davidson, ultimately conceded that a purely coherent web of belief risks epistemic isolation unless constrained by experiential or causal contact with the world. Hybrid or "constraint" models now dominate: Susan Haack's *foundherentism* combines internal coherence with empirical grounding (Haack 1993); Hilary Putnam's *internal realism* ties truth to both rational coherence and responsiveness to reality (Putnam 1981); and contemporary truth-maker theorists (e.g., Armstrong 2004; Mulligan et al. 1984) maintain that every true proposition must correspond to some fact or state of affairs that makes it true. In short, the prevailing consensus is that while coherence secures the rational unity of belief, truth still requires correspondence or grounding—some form of answerability to the world that resists and corrects even the most consistent fictions. This prevailing view sets a clear philosophical backdrop for understanding LLM hallucinations: Their textual coherence may simulate truth, but without external grounding, their outputs remain epistemically suspended—plausible, yet unanchored in reality.

However, in my framing, the hallucination problem diverges from the coherence–correspondence debate in two important respects. (1) First, coherence theories of truth rely on the representation of propositions and the ability to assess their coherence, i.e., their logical compatibility. But as argued above, LLMs have limited reasoning capacities precisely because they are incapable of representing propositions. Therefore, even if perfect coherence were sufficient—without the need for grounding in corresponding facts—LLMs would still be unable to achieve it. The coherence that LLMs exhibit operates at the level of tokens rather than the level of propositions. Thus, although LLMs are based on a holistic account of meaning, it is word or token holism, not propositional holism. This crucial feature of LLMs' architecture will be of utmost importance in the following discussion. (2) The coherence–correspondence debate pertains to an epistemological issue, while the

 doi:10.20944/preprints202601.1447.v1

hallucination problem is also a structural issue of LLM architecture. This is to say, hallucinations arise due both to the truth source problem—which is an epistemological issue—and, primarily, the truth representation problem—which is a structural issue.

III. Implicit Theory of Language

LLMs operate on what can be described as an implicit theory of language grounded in *vector semantics*, also known as distributional semantics. In this framework, meaning is not defined by reference to external objects or truth conditions, but by the statistical relations among linguistic items across vast corpora of usage. Each word, phrase, or sentence is represented as a high-dimensional vector in a learned semantic space, and meaning arises from geometric proximity and relational structure: Words that occur in similar contexts acquire similar vector representations. Unlike earlier models such as word2vec and GloVe, which produced static embeddings, transformer-based LLMs create contextualized embeddings whose values depend on surrounding words, allowing for dynamic shifts in meaning across linguistic environments. Consequently, language is modeled as a continuous, predictive geometry of associations rather than as a compositional, truth-conditional, or rule-based system. This internal structure implies an empiricist and associationist view of language, in which understanding is measured by statistical prediction and pattern recognition rather than by shared human practices, intentionality, or referential grounding—something many critics have described as *ungrounded* or *disembodied* semantics (Harnad 1990; Weatherby 2025; Felin and Holweg 2024; Šekrst forthcoming).

As convincingly argued by Leif Weatherby in his recent book *Language Machines*, LLMs can be seen as the most radical technological realization of Saussure's vision of language as a self-contained system of differences without positive terms. Saussure held that meaning arises not from reference to external things but from the network of internal relations among signs within the *langue*. LLMs reproduce precisely this structural condition: Their entire semantic universe is generated from the statistical interplay of linguistic signs with no access to extralinguistic reality. Each token acquires its significance only through differential positioning within the model's vast vector space, echoing Saussure's notion that "in language there are only differences, without positive terms." (Saussure 1983, 166) The model's training on patterns of co-occurrence operationalizes the *paradigmatic* and *syntagmatic* relations that, for Saussure, define meaning within the linguistic system. Thus, LLMs instantiate a purely immanent and relational semantics—a computational *langue* detached from referential anchoring—and so fulfill, in algorithmic form, what Weatherby describes as the Saussurean dream of language without referent.

I have arrived at a crucial point in my argument where I can formulate the following aporetic claims: LLMs' *success* can be explained due to their reliance on the structuralist account of language, which can get along without external reference. Conversely, LLMs' *malfunction* can be attributed to their reliance on this structuralist account of language. As aporetic as these claims may seem, they do not exclude each other, as they can both be true. However, they misplace the issue because they focus on linguistic signs and their reference. As suggested above, the reason why LLMs hallucinate is their inability to properly represent propositions, i.e., the truth representation problem.

In AI research, the problem of reference is known as the *symbol grounding problem*, a concept introduced by Harnad in 1990. It would be instrumental for my argument to examine the recent reframing of this problem in the context of LLMs by Mollo and Millière (2025) under the label of the *vector grounding problem*. The issue is this: Can purely relational semantics be supplemented by grounding in real-world entities? Symbol grounding refers to the concept of reference. In LLMs, symbols are converted into tokens and embedded in vectors. These embeddings are entirely relational. However, they establish causal-historical connections to real-world entities through the training process. Hence, for instance, the embedding for a proper name (say, "Elon Musk") is influenced by the occurrence of the string "Elon Musk" in the training dataset, which, in turn, is influenced by the real person Elon Musk. The issue with a causal account of reference is its vagueness

 doi:10.20944/preprints202601.1447.v1

and the difficulty in retracing its steps. The relevant causal links can easily be disrupted, a concern I refer to as the truth source problem.

In the face of these problems, Mollo and Millière suggest that a model's learning can establish "extra-linguistic, world-involving functions." They write:

> This is most evident in models that undergo fine-tuning to satisfy human preferences, since such fine-tuning explicitly involves selecting internal states that increase the probability of outputs satisfying world-involving norms, such as factual accuracy. However, we will also present evidence suggesting that pre-training alone can, in some contexts, select for internal states with world-involving content, albeit in a more indirect way. (Mollo and Millière 2025, 17)

This is in line with what was established above: The aim of fine-tuning is to combat hallucinations, that is, to enhance factual accuracy. However, this solution overlooks the numerous issues associated with fine-tuning, such as representation drift.

The symbol/vector grounding problem marks an attempt to supplement the relational semantics with a robust notion of symbol reference. In my view, this attempt is futile. Our language is full of nonreferential terms (or, alternatively, the only properly referential terms are proper names). And even if we managed to ground every symbol/vector, the problem with hallucinations would persist. An LLM could recombine properly grounded terms in a way that the output would express a false proposition. On the other hand, we can admit that there is a grain of truth in the grounding problem. However, what needs to be grounded are not individual symbols, but rather propositions (ideally embedded in vectors). To implement this, we will have to abandon the structuralist account of language in favor of one that takes propositions, not symbols, as the primary units of linguistic meaning. As we shall see in the next section, such a program has recently been suggested by Chalmers.

IV. Chalmers's Propositional Interpretability and Davidson's Radical Interpretation

In his recent paper "Propositional Interpretability in Artificial Intelligence," which is still a "rough draft," Chalmers discusses many insights that overlap with those presented in the present article. It is essential to clearly indicate where our views converge and diverge.

Chalmers argues that to understand AI systems, we should interpret their mechanisms through the lens of generalized propositional attitudes. Propositional attitudes describe a subject's relation to a proposition and include beliefs, desires, and assessments of probability. For example, we might say, "x believes that P," "x desires that P," or "x thinks that P is probable." Chalmers interprets these attitudes quite broadly; even representing a proposition for a specific purpose or goal can qualify as a propositional attitude. These attitudes connect the thinking mind (the subject) to propositions represented by that subject. Chalmers convincingly asserts that attributing propositional attitudes to a system does not entail that the system is conscious or possesses mental states. This perspective allows him to explore whether AI systems can represent propositional attitudes without delving into the debate over whether LLMs are conscious. He goes on to argue that to truly understand a system, including but not limited to LLMs, it is not enough to be clear about how the system represents propositions. Rather, it is crucial to understand the system's attitude toward these propositions (specifically, what the system aims to do with them and how it intends to utilize them).

Chalmers distinguishes between conceptual interpretability and propositional interpretability, arguing that the former is insufficient for understanding AI systems. Conceptual interpretability takes concepts as the primary unit of representation. Although Chalmers points out several weaknesses of conceptual interpretability, one is crucial for our present discussion: Conceptual interpretability fails to represent how concepts are combined into propositions. A mere cluster of

 doi:10.20944/preprints202601.1447.v1

concepts is not enough.[7] This issue suggests that mere conceptual interpretability is insufficient to represent the truth accurately. Chalmers speaks of "weaknesses [that] arise from fragility and ground truth." Another point highlighted by Chalmers is that current LLMs are capable of conceptual interpretability, as they represent conceptual tokens as high-dimensional vectors. This echoes the idea, discussed above, that LLMs embody a structuralist account of language, which is a kind of conceptual interpretability.

So far, I agree with Chalmers's line of argument. One key question at this stage is whether LLMs are capable of propositional interpretability. A direct answer would be that they are not, due to numerous persistent issues (mentioned by Chalmers as well as earlier in the present paper). Here, we are still in agreement. However, an even more pertinent question would be: Can current LLMs with their inherent vector semantics be refined to better represent propositions and propositional attitudes? Can propositional interpretability be built out of conceptual interpretability? This is where our disagreement begins, as Chalmers is positive that this can be done, while I think that conceptual interpretability is an unsuitable starting point. One should begin with propositional interpretability, that is, with basic representations of propositions and attitudes toward them. This may come down to a technical question.

However, there is a deep underlying philosophical problem, which can be formulated without reference to AI. Are concepts linguistic primitives, that is, basic building blocks of the whole language, from which propositions can be constructed? Or are propositions semantically primary, such that names gain meaning through their place in them? The last clause expresses the *context principle*, advocated in various forms by Frege and Wittgenstein.[8] Among advocates of the primacy of simple concepts are the early Russell and the early Carnap. Carnap, in *The Logical Structure of the World* (1928), attempts to build knowledge compositionally from atomic "elementary experiences". His project exemplifies the ambition to begin with atoms and construct propositions from them.

The context principle was radicalized into full-blown holism by Quine and Davidson, who argued that the meaning of a proposition can be determined only within the context of the whole theory or the whole language.[9] Meaning and verification are holistic, not atomic: No statement (let alone a single concept) is meaningful in isolation. Quine thus extends Frege's context principle to the level of theory. Although Davidson subscribed to Quine's holism, in "Truth and Meaning" (1967) he argues that an empirical semantic theory must begin with sentences as truth-bearing entities. Words have meaning only within a truth-theoretic structure based on whole propositions. Davidson imagines a language user who, so to speak, builds their empirical semantic theory from scratch, without presupposing any fixed word or sentence meaning—a process that he calls radical interpretation.

Let us turn back to Chalmers. His *Constructing the World* (2012) lays out a neo-Carnapian project that attempts a conceptual reconstruction of knowledge in modal-semantic terms. However, Chalmers does not follow Carnap's logical atomism of concepts; he replaces it with a propositional and modal framework. His recent draft paper on propositional interpretability is Carnapian in essence. In that paper, Chalmers sets out the view that propositional interpretability can be constructed out of conceptual interpretability.

To position propositional interpretability, Chalmers invokes the distinction between the basic/ground layer and middle layer in LLMs' design. The basic layer encodes token meanings as high-dimensional vectors. In the current architecture, a single vector cannot represent a whole

---

[7] This echoes Russell's prolonged struggles with the problem of the unity of propositions.

[8] Frege insists that "never is the question to be raised about the meaning of a word in isolation, but only in the context of a sentence" (Frege 1950, §62). Wittgenstein develops a structurally similar view when he states that "only the proposition has sense; only in the context of a proposition has a name meaning" (Wittgenstein 1922, 3.3).

[9] Quine: "The unit of empirical significance is the whole of science" (1951, 42).

 doi:10.20944/preprints202601.1447.v1

proposition, and especially not a propositional attitude. The middle layer(s) serve(s) as a bridge from lexical/syntactic encodings (in the lower layers) to the very high-level tasks or generation encodings close to the generation of output probabilities (in the top layers). Based on a study by Meng et al. (2023), which I referred to earlier, Chalmers argues that the middle layer is capable of storing and computing factual associations, and thus it can be vital to representing propositions (this idea dates back to fact-checking strategies in the 1970s; we shall return to this topic later).

The crucial point is that the middle layer is constituted during post-training by various optimization and model-editing techniques. As argued above, attempts to represent propositions in this layer are beset by various problems and flaws. Although Chalmers is aware of these issues, he appears to see them as bugs to be improved. He is confident that representing propositions in the middle layer constitutes a form of propositional interpretability. I, in contrast, believe that attempting to represent propositions in the middle layer is a fundamental architectural flaw in the design of LLMs. Let us look at Chalmers's argument in more detail. He writes:

> A related objection is that current language models lack beliefs because they do not value truth: they have been trained only to predict the next word, not to say what is true. Now, as many have observed in response, current language models typically undergo a round of fine-tuning by reinforcement learning, where true answers are rewarded. Even in the absence of explicit training, it may well be that optimal performance in predicting the next word requires having generally true beliefs about the world. Either way, truth may be rewarded in the training process, albeitly imperfectly in a way that leaves room for much unreliability. (Chalmers 2025, 24)

Let us unpack Chalmers's argument. He appears to argue that optimal performance in predicting the next word requires genuine beliefs about the world. But optimal in which sense? If "optimal" means "statistically most probable"—which is how current LLMs are designed—then no beliefs about the world are required. Factual accuracy is only an unreliable by-product of statistical optimality. If "optimal" means "truthful," that is, with no or minimal hallucinations, then there is a necessary clash between statistical and veridical optimality. As already pointed out, there are claims that are statistically less probable but still true. This is to say that statistical optimality on the level of words/tokens is something different from veridical optimality on the level of propositions and facts. Chalmers seems to be not entirely sensitive to this distinction.

His disregard of the distinction is obvious when he writes: "But I think language models can engage in at least structural representation of the non-linguistic world. Representation of the world is made easier by the fact that language models already use natural language, and arguably inherit their meanings" (Chalmers 2025, 24). Here, we must distinguish representation of meaning, which LLMs are capable of (in the ground/conceptual layer), from representation of the nonlinguistic world, that is, of true factual knowledge. LLMs extract linguistic meaning from their training data. I wish to insist, contra Chalmers, that they do not inherit reliable true knowledge (the truth source problem) and are not capable of representing such knowledge (the truth representation problem).

Therefore, it appears that constructing propositional interpretability from conceptual interpretability cannot yield a reliable representation of a proposition. This is, in my view, the main cause of LLMs' hallucinations.

There is a possible way out of this problem (briefly suggested earlier) that is available to Chalmers. If propositions are in essence pictorial—something along the lines suggested in Wittgenstein's *Tractatus* or in contemporary accounts of propositions inspired by the *Tractatus*[10]—

[10] Dominic Gregory (2020) argues that pictorial representations possess a logical form akin to predication—depicting things as having properties—and that, when contextually framed, such images can bear propositional content. Tue Trinh (2024) likewise treats propositions as pictorial—not in a visual sense, but a structural one: They "show" possible states of affairs through their internal projective form. Both authors thus recast the

 doi:10.20944/preprints202601.1447.v1

then the structure of a proposition reflects the structure of the nonlinguistic fact it represents. The structural correspondence between a proposition and the fact it represents may be complex. However, in a similar manner, we can say that there can be a complex correspondence between the structure of the proposition and the statistical patterns computed by an LLM. In this sense, the model could capture the structural representation of the nonlinguistic world as suggested by Chalmers. This promising line of inquiry must be set aside for future work.

In the remaining part of this section, I shall turn to another philosophical underpinning of Chalmers's notion of propositional interpretability, which is, perhaps paradoxically, Davidson's radical interpretation (or the variant of it suggested by Lewis).

Radical Interpretation

Chalmers is correct that radical interpretation is a crucial philosophical framework that can be instrumental in understanding LLMs. He is also correct that we should follow Lewis's version of radical interpretation, which dispenses with Davidson's behaviorism and allows us to consider internal physical states of the system.[11] Hence, given a set of facts about what the system says, *P* (output), together with the facts about the physical system, we have to find out what the system means by *P* and its attitude toward *P* (propositional meaning and propositional attitude). Radical interpretation treats meaning and attitudes (beliefs) holistically. One cannot (and does not need to) separate what the language user means from what they believe about what they mean. In other words, propositions are not separable from attitudes. All this supports Chalmers's approach. We do not need separate representations of propositions and attitudes toward them.

As should be clear from the discussion so far, hallucinations emerge because LLMs are not sufficiently capable of representing truth. At the same time, truth is, for Davidson, the most central notion. This centrality is captured by his *principle of charity*. However, Chalmers does not address this principle in his appropriation of Davidson's radical interpretation. This principle serves as a methodological guideline for radical interpretation. Simplifying somewhat, the principle of charity suggests that the radical interpreter should assume that the system being analyzed (in this case, a language model) has a coherent set of beliefs and that most of what the system expresses is intended sincerely, that is, is *true*. Davidson refers to these two aspects of the principle of charity as the principle of coherence and the principle of correspondence,[12] echoing the coherence–correspondence debate mentioned earlier. This is not to say that the interpreter must automatically assume that everything that others say is always true and coherent. That would be naïve. However, any defect (mistakes, errors, lies, rational incoherencies, etc.) can be detected only against a background of predominant coherence and truthfulness. As Davidson says, "disagreement and agreement alike are intelligible only against a background of massive agreement" (Davidson 1984, 137). If a language user

---

proposition as a picture in the formal, not perceptual, sense, whose meaning arises from the way its internal configuration mirrors the world's possible arrangements.

[11] Lewis's approach is suitable for interpreting artificial machines, because their inner architecture and workings are completely transparent to us—which cannot be said about the human mind.

[12] "The process of separating meaning and opinion invokes two key principles which must be applicable if a speaker is interpretable: the Principle of Coherence and the Principle of Correspondence. The Principle of Coherence prompts the interpreter to discover a degree of logical consistency in the thought of the speaker; the Principle of Correspondence prompts the interpreter to take the speaker to be responding to the same features of the world that he (the interpreter) would be responding to under similar circumstances. Both principles can be (and have been) called principles of charity: one principle endows the speaker with a modicum of logic, the other endows him with a degree of what the interpreter takes to be true belief about the world" (Davidson 2001, 211).

 doi:10.20944/preprints202601.1447.v1

(human or artificial) were to display massive incoherence or falsity in their statements, we would not consider them a rational agent.[13]

My claim now is that LLMs display such massive falsity in hallucinating. I wish to insist that hallucinations are *not* an occasional phenomenon on par with human lies or factual mistakes. The ubiquity of hallucinations renders the principle of charity untenable. Hence, we are forced to conclude that LLMs cannot be interpreted as rational agents in Davidson's framework.

To avert this conclusion, Chalmers must downplay the problem of hallucination. He discusses the following objection: "A common objection is that current AI systems don't have beliefs because they're too unreliable. They famously give wrong answers to many questions." To which he replies: "On the other hand, humans give many wrong answers too, and it's not obvious why this should undermine beliefs entirely, however. There are many issues on which current AI systems give consistently correct answers, suggesting true beliefs." This is to say that there is no significant difference between humans' occasional mistakes and lies and LLMs' inclination to hallucinations—that is, their tendency to produce factually false content. I hope to have made the case that LLMs' hallucinations are not random errors, but rather point to a structural problem in LLMs' design.

V. Toward the Proposed Solution: Atomic Facts/Propositions in the Basic Layer

The idea that propositions are represented in the middle (symbolic/reasoning) layer, rather than in the basic layer of raw parameters or vectors, is not new. Various fact-checking strategies operate in this layer. However, these strategies endorse the idea of atomic propositions or atomic facts, where knowledge is explicitly encoded as the smallest indivisible, independently verifiable propositions—typically expressed as subject–predicate–object triples or *n*-ary relations (Russell and Norvig 2010; Brachman and Levesque 2004). In the LLM era, researchers decompose model outputs into such atoms, verify each separately, and reassemble only supported claims, yielding major gains in factual accuracy (Min et al. 2023; Zheng et al. 2025). Grokipedia (xAI, launched October 2025) is the first production system to make this the core of an entire encyclopedia. Early independent analyses (Tuquero 2025) indicate that Grokipedia still inherits source-selection biases from Grok's training data and occasionally produces confidence-inflated scores on politically charged topics where atomic decomposition alone cannot fully resolve interpretive nuances. In short, the problem of hallucination persists.

If my argument above is correct, the culprit is the idea that propositions are represented in the middle layer. On the other hand, the idea of *atomic propositions* points to the solution I propose. The motivation is obvious: We must reduce the number of propositions to be represented in the language model, ideally to a finite number within technical capabilities. If today's LLM tokens are atomic units of word-meaning, then atomic propositions are units of proposition-meaning. This brings us to the key question: Can atomic propositions be represented in the basic layer?

Representing atomic facts/propositions directly in the base parameters of LLMs—rather than as emergent by-products of token prediction—remains an unsolved challenge in 2025. Early small-scale attempts trained transformers explicitly on knowledge graph triples (Yao et al. 2019), but factual knowledge in today's billion-parameter models is still massively distributed and polysemantic, with individual neurons participating in thousands of unrelated concepts (Elhage et al. 2022;Li et al. 2022). Probing studies have identified sparse, near-monosemantic directions in activation space associated with truth-conditional or factuality-related content (Zou et al. 2023; Bao et al. 2025), and dictionary-learning approaches have begun extracting millions of interpretable features that in some cases correspond to proposition-like or atomically structured semantic content (Bricken et al. 2023;

[13] Claudine Verheggen and Robert Myers argue in a recent article (2025) that the principle of charity is the essential ingredient of meaning itself in all contexts. Therefore, in order to get their interpretation off the ground, the radical interpreter must assume this principle. It is not just a methodological maxim, but an integral component of their broader approach. The principle of charity cannot be excluded from radical interpretation.

 doi:10.20944/preprints202601.1447.v1

Templeton et al. 2024). Yet attempts to scale these discoveries into a systematic, native basic-layer representation are still confounded by representation entanglement, catastrophic forgetting during updates, and the absence of training objectives that reliably induce clean atomic factorization at scale. As of late 2025, no production LLM encodes atomic facts directly in its base weights; all current factual reliability still depends on middle-layer overlays.

Crucially, atomic propositions are indeed represented in the basic layer. However, the representation of a single proposition is distributed over a vast number of vectors. The possibility of representing individual atomic propositions as single, dedicated vectors in the base parameters of LLMs remains unrealized in production systems as of late 2025. Current transformer-based LLMs encode facts in distributed, polysemantic form via superposition (Olah et al. 2020; Elhage et al. 2022). The closest approximation is provided by sparse autoencoders (SAEs), which decompose activations into millions of sparse, monosemantic features—each corresponding to a high-dimensional direction that activates primarily for one interpretable concept or proposition, achieving 70–90% human-labeled interpretability in recent models (Bricken et al. 2023; Cunningham et al. 2024; Templeton 2024). These features support causal interventions (Meng et al. 2023) but remain post hoc reconstructions rather than native parameter-level representations. Complementing this, large concept models (LCMs) advance higher-level atomic representations by operating directly in a semantic embedding space of "concepts"—abstract atomic ideas akin to sentences—using SONAR embeddings for multilingual, multimodal one-to-one mappings (Barrault et al. 2024). However, fully native, end-to-end, one-proposition–one-vector encoding remains an open research challenge (Sharkey et al. 2025).

There is a significant philosophical underpinning to the idea of atomic facts or atomic propositions, which emerged in logical atomism at the beginning of the twentieth century. Its seminal expressions can be found in Wittgenstein's *Tractatus Logico-Philosophicus* (1922) and Russell's *Philosophy of Logical Atomism* (1918). I give preference to Wittgenstein's *Tractatus* because he clearly articulates the primacy of propositions over words, as already alluded to earlier in the discussion of the context principle. Atoms are usually the smallest, not further divisible units, as Russell seems to suggest.[14] However, atomic propositions and facts are divisible. The defining characteristic that makes them atomic is their mutual *independence*.[15] Hence, a proposition is atomic not because it is indivisible but rather because its truth (or falsity) is independent of other atomic propositions. I shall now examine how this requirement of independence has been implemented in LLMs.

The main issue with representing propositions in the middle layer is that their representations are not sufficiently independent. There are what are known as "representation entanglements": Modifying one proposition can impact others and ultimately distort the entire model. In my view, propositions should only be represented in the basic layer to ensure that they remain mutually independent. Such independence should concern only those propositions that are independent in reality. That is why I shall restrict my focus to atomic propositions. However, even here, as we have seen above, some representations are distributed across many vectors, which does not resolve the problem of representation entanglement. LLMs are based on multidimensional linear spaces where each vector is mutually independent of any other. It would therefore be natural for each atomic proposition to be represented by one vector.

Many of the technical proposals or attempted implementations mentioned above fulfill these demands only partially. The relevant papers—particularly those on SAEs and dictionary learning—consistently define "atomic" as referring to the simplest, indivisible units of representation that capture a single, discrete proposition without further decomposition. While independence emerges

[14] Russell writes: "The simplest imaginable facts are those which consist in the possession of a quality by some particular thing. […] The whole lot of them, taken together, are as facts go very simple, and are what I call atomic facts. The propositions expressing them are what I call *atomic propositions*" (Russell 1918/2010, 26–27).

[15] Thus, Wittgenstein writes: "Atomic facts are independent of one another" (1922, 2.061). And: "From an atomic proposition no other can be inferred" (ibid., 5.134).

 doi:10.20944/preprints202601.1447.v1

as a by-product (facilitating modular composition), the primary emphasis across these works is on simplicity and indivisibility in order to achieve faithful, interpretable basic-layer encodings. In LLMs, the concept "atomic" is explicitly tied to the definition of a concept as "an abstract atomic idea," with atomicity emphasizing indivisibility and simplicity as the minimal semantic unit (Barrault et al. 2024). However, Barrault et al. strongly emphasize explicit *independence* as a key property in the design of LCMs. The emphasis on independence underscores practical benefits for hierarchical reasoning and multilingual/multimodal scalability, distinguishing LCMs from token-level models.

As promising as these proposals might be, they face various technical, architectural, and philosophical challenges. One architectural issue is the exponential growth in the number of vectors that represent atomic propositions, which far exceeds the approximately 50,000 linguistic tokens currently handled by LLM architecture. Another architectural and philosophical issue is how to recognize atomic propositions if simplicity is not their defining characteristic. Wittgenstein (1929) himself came to realize that simple, not further analyzable propositions expressing qualitative degrees are not mutually independent (one proposition can exclude another). This might indicate, Wittgenstein concluded, that we have not found the ultimate analysis of these propositions. However, shortly after that, he abandoned the search for such an analysis and proclaimed the idea of atomic propositions to be erroneous (1980, 119).

LLMs, with their inherent vector semantics and ability to accommodate billions of parameters, give us hope that this search for an acceptable analysis does not need to be abandoned. This development could have significant implications for the revival of the program of logical atomism. While current LLMs reflect a perspective on language derived from Saussurean structuralism, future LLMs capable of dealing with the hallucination problem could be based on the principles of logical atomism.

VI. Conclusions

Throughout the article, I have developed and defended the thesis that LLMs' hallucinations arise from their inability to accurately represent propositions that can be true or false. I presupposed that hallucinations are not a marginal phenomenon that can be tolerated; rather, they present a serious issue that degrades the performance and applicability of LLMs. I identified two interrelated yet distinct issues: the truth source problem and the truth representation problem. The former is an epistemological issue relating to the training data source, the latter a structural flaw in the design of LLMs. Regarding the latter, LLMs lack an adequate representation of propositions, due to a combination of technical and philosophical reasons. It turns out that the makeup of LLMs is based on an implicit account of language that does not give sufficient weight to propositions. To clarify this issue, I have presented three underlying accounts of language and three ways of representing propositions in LLMs.

(1) The first account concerns how LLMs currently operate. In this model, concepts/tokens are embedded in vectors that express their position in a high-dimensional vector space. A token's meaning is determined solely by its relation to other tokens within this vector space, with no need for external references. A proposition is then represented as a series of such embeddings. As Weatherby (2025) recently argued, this account of linguistic signs conforms to classical Saussurean structuralism. He further argues that LLMs' primary function is to capture and express the *general poetics* of language (as conceived by Roman Jacobson) and not to produce factually accurate statements. There are, no doubt, useful employments of LLMs in this mode, such as linguistic pattern recognition. Hallucinations are not a (major) problem from this perspective. However, I have approached matters from a different point of departure, where they *do* pose a problem, and so I have sought ways to address it.

(2) The second account of how LLMs operate builds on the first. The main difference is that it directly addresses the problem of representing propositions. Just as tokens are represented by vectors in the basic layer, propositions are represented in the middle layer, which builds upon the basic layer. The representation of a single proposition is distributed across many vectors, including those that

 doi:10.20944/preprints202601.1447.v1

represent tokens not directly involved in the proposition. Following the recent work of Chalmers, the underlying account of language is that of Davidson's radical interpretation. However, as I have argued, representing propositions in the middle layer faces crucial issues, most notably that of representation drift. With the problem of hallucinations unresolved, embracing Davidson's radical interpretation is challenging because its key premise, the principle of charity, becomes untenable.

(3) The third account also recognizes the necessity of representing propositions. However, this time it is claimed that propositions should be represented in the basic layer of vectors/model weights. In the ideal case, one vector would correspond to one proposition. As there is a finite number of vectors, such systems would be able to represent a finite number of propositions, which I call "atomic propositions." The underlaying view of language is that of logical atomism. However, here we must be more specific and focus on a kind of logical atomism that takes atomic propositions as mutually independent logical atoms. As it turns out, this is the view espoused in Wittgenstein's *Tractatus*.

Having completed my argument, I can now reveal the continuation of the discussion between Pinker and Dawkins introduced at the beginning of my essay. Dawkins raises an intuitive objection to Pinker's discussion of hallucinations:

> Why couldn't you just look it up in Google?

Pinker replies with the following informed answer:

> So there are now hybrid models that will, before they produce the output, they'll kind of look them up on Google. And not surprisingly, Google itself. I mean, that was what Gemini was originally designed to do. (Dawkins and Pinker 2025, 1:04:25–42)

The argument presented here has sought to demonstrate, in line with Pinker's original suggestion, that such hybrid models fail to address the truth representation problem and, consequently, are ineffective at dealing with hallucinations.

Viewed more broadly, the current argument suggests that different designs of LLMs can be seen as practical implementations of two main philosophical approaches to language. One approach focuses on concepts, exemplified by classical structuralism and Russell's logical atomism, while the other centers on propositions, drawing from Wittgenstein's atomism and Davidson's radical interpretation. The issue of hallucinations in LLMs, and how they are addressed, serves as a touchstone for these philosophical frameworks. The success of current LLMs is also a testament to the structuralist account of language; conversely, the flaws in the current makeup of LLMs, notably hallucinations, must be attributed to inherent flaws in the structuralist program. At this time, it is not possible to draw equivalent conclusions regarding logical atomism or radical interpretation, as there are not yet any practical implementations of LLMs that embody these views of language.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript." Please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: Please add: "This research received no external funding" or "This research was funded by NAME OF FUNDER, grant number XXX" and "The APC was funded by XXX". Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding. Any errors may affect your future funding.

Institutional Review Board Statement: In this section, please add the Institutional Review Board Statement and approval number for studies involving humans or animals. You might choose to exclude this statement if the

 doi:10.20944/preprints202601.1447.v1

study did not require ethical approval. Please note that the Editorial Office might ask you for further information. Please add "The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving humans. OR "The animal study protocol was approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving animals. OR "Ethical review and approval were waived for this study due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving humans or animals.

Informed Consent Statement: Any research article describing a study involving humans should contain this statement. Please add "Informed consent was obtained from all subjects involved in the study." OR "Patient consent was waived due to REASON (please provide a detailed justification)." OR "Not applicable." for studies not involving humans. You might also choose to exclude this statement if the study did not involve humans. Written informed consent for publication must be obtained from participating patients who can be identified (including by the patients themselves). Please state "Written informed consent has been obtained from the patient(s) to publish this paper" if applicable.

Data Availability Statement: We encourage all authors of articles published in MDPI journals to share their research data. In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Where no new data were created, or where data is unavailable due to privacy or ethical restrictions, a statement is still required. Suggested Data Availability Statements are available in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics.

Acknowledgments: In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments). Where GenAI has been used for purposes such as generating text, data, or graphics, or for study design, data collection, analysis, or interpretation of data, please add "During the preparation of this manuscript/study, the author(s) used [tool name, version information] for the purposes of [description of use]. The authors have reviewed and edited the output and take full responsibility for the content of this publication.".

Conflicts of Interest: Declare conflicts of interest or state "The authors declare no conflicts of interest." Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results must be declared in this section. If there is no role, please state "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results".

References

Armstrong, David M. 2004. *Truth and truthmakers*. Cambridge: Cambridge University Press.

Bao, Yuntai; Zhang, Xuhong; Du, Tianyu; Zhao, Xinkui; Feng, Zhengwen; Peng, Hao; Yin, Jianwei. 2025. Probing the geometry of truth: Consistency and generalization of truth directions in LLMs across logical transformations and question answering tasks. *arXiv preprint arXiv*:2506.00823. Available online: https://arxiv.org/abs/2506.00823 (accessed December 25, 2025).

Barrault, Loïc; Duquenne, Paul-Ambroise; Elbayad, Maha; et al. 2024. Large concept models: Language modeling in a sentence representation space. *arXiv preprint* arXiv:2412.08821. Available online: https://arxiv.org/abs/2412.08821 (accessed December 19, 2025).

Brachman, Ronald J.; Levesque, Hector J. 2004. *Knowledge representation and reasoning*. San Francisco: Morgan Kaufmann.

Bricken, Trenton; Templeton, Adly; Batson, Joshua; et al. 2023. Towards monosemanticity: Decomposing language model activations with dictionary learning. *Transformer Circuits*. Available online: https://transformer-circuits.pub/2023/monosemantic-features (accessed December 25, 2025).

 doi:10.20944/preprints202601.1447.v1

Broniatowski, David A.; Jamison, Amelia M.; Qi, SiHua; AlKulaib, Lulwah; Chen, Tao; Benton, Adrian; Quinn, Sandra C.; Dredze, Mark. 2018. Vaccine discourse in the era of social media: Vaccine denialism, misinformation, and trust. *American Journal of Public Health* 108 (S2), S150–S157. https://doi.org/10.2105/AJPH.2018.304567.

Carnap, Rudolf. 1928. *Der logische Aufbau der Welt*. Berlin: Weltkreis-Verlag.

Chalmers, David J. 2012. *Constructing the world*. Oxford: Oxford University Press.

Chalmers, David J. 2025. Propositional interpretability in artificial intelligence. *arXiv preprint* arXiv:2501.15740. Available online: https://arxiv.org/abs/2501.15740 (accessed December 19, 2025).

Chang, Haw-Shiuan; McCallum, Andrew. 2022. Softmax bottleneck makes language models unable to represent multi-mode word distributions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022): Long Papers*; Association for Computational Linguistics: Dublin, Ireland, 8048–8073.

Chekalina, Veronika; Kutuzov, Andrey; Anjos, André. 2024. Addressing hallucinations in language models with knowledge graph embeddings as an additional modality. *arXiv preprint* arXiv:2411.11531. Available online: https://arxiv.org/abs/2411.11531 (accessed December 19, 2025).

Chen, Shiqi; Xiong, Miao; Liu, Junteng; et al. 2024. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. In *Proceedings of the 41st International Conference on Machine Learning*; PMLR 235, 7553–7567. Available online: https://arxiv.org/abs/2403.01548 (accessed December 25, 2025).

Cossio, Manuel. 2025. A comprehensive taxonomy of hallucinations in large language models. *arXiv preprint* arXiv:2508.01781. Available online: https://arxiv.org/abs/2508.01781 (accessed December 19, 2025).

Cunningham, Hoagy; et al. 2024. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint* arXiv:2309.08600. Available online: https://arxiv.org/abs/2309.08600 (accessed December 25, 2025).

Davidson, Donald. 1967/1984. Truth and meaning. In *Inquiries into truth and interpretation*. Oxford: Clarendon Press, 17–36.

Davidson, Donald. 1984. Radical interpretation. In *Inquiries into truth and interpretation*. Oxford: Clarendon Press, 125–139.

Davidson, Donald. 1986. A coherence theory of truth and knowledge. In *Truth and interpretation*, ed. Ernest LePore. Oxford: Blackwell, 307–319.

Davidson, Donald. 2001. Three varieties of knowledge. In *Subjective, intersubjective, objective*. Oxford: Clarendon Press, 205–220.

Dawkins, Richard; Pinker, Steven. 2025. Can we still be optimistic about the future? A conversation with Steven Pinker. YouTube video on *The Poetry of Reality with Richard Dawkins*, published January 15, 2025. Available online: https://www.youtube.com/watch?v=qFZ8_Ide-aA (accessed December 19, 2025).

De Cao, Nicola; Aziz, Wilker; Titov, Ivan. 2021. Editing factual knowledge in language models. In *Findings of EMNLP 2021*; Association for Computational Linguistics, 1649–1660. Available online: https://aclanthology.org/2021.emnlp-main.522.pdf (accessed December 19, 2025).

Elhage, Nelson; et al. 2022. *Toy models of superposition*. Transformer Circuits Thread (online PDF). Available online: https://transformer-circuits.pub/2022/toy_model/toy_model.pdf (accessed December 25, 2025).

Felin, Teppo; Holweg, Matthias. 2024. Theory is all you need: AI, human cognition, and causal reasoning. *Strategy Science* 9 (4), 346–371. https://doi.org/10.1287/stsc.2024.0189 (accessed December 25, 2025).

Frankfurt, Harry G. 2005. *On bullshit*. Princeton: Princeton University Press.

Frege, Gottlob. 1884. *Die Grundlagen der Arithmetik*. Breslau: Wilhelm Koebner.

Frege, Gottlob. 1950. *The foundations of arithmetic*. Oxford: Blackwell.

Gekhman, Dor; Schoelkopf, Hailey; Geva, Mor; Goldberg, Yoav. 2024. *Does fine-tuning LLMs on new knowledge encourage hallucinations? arXiv preprint* arXiv:2405.05904. Available online: https://arxiv.org/abs/2405.05904 (accessed December 25, 2025).

Ghosal, Gaurav; Hashimoto, Tatsunori; Raghunathan, Aditi. 2024. Understanding finetuning for factual knowledge extraction. *arXiv preprint* arXiv:2406.14785. Available online: https://arxiv.org/abs/2406.14785 (accessed December 25, 2025).

Gregory, Dominic. 2020. Pictures, propositions, and predicates. *Philosophical Studies* 177, 1567–1588.

 doi:10.20944/preprints202601.1447.v1

Haack, Susan. 1993. *Evidence and inquiry*. Oxford: Blackwell.

Harnad, Stevan. 1990. The symbol grounding problem. *Physica D* 42, 335–346. https://doi.org/10.1016/0167-2789(90)90087-6.

Huang, Lei; Yu, Weijiang; Ma, Weitao; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43, 1–55. https://doi.org/10.1145/3703155.

Jang, Joel; Ye, Seonghyeon; Lee, Changho; et al. 2022. Towards continual knowledge learning of language models. *arXiv preprint* arXiv:2110.03215. Available online: https://arxiv.org/abs/2110.03215 (accessed December 19, 2025).

Jolley, Daniel; Douglas, Karen M. 2014. The effects of anti-vaccine conspiracy theories on vaccination intentions. *PLOS ONE* 9 (2), e89177. https://doi.org/10.1371/journal.pone.0089177.

Joshi, Satyadhar. 2025. Mitigating LLM hallucinations: A comprehensive review of techniques and architectures. *Preprints*. Available online: https://www.preprints.org/manuscript/202505.1955/v1 (accessed December 19, 2025).

Lavrinovics, Ernests; Biswas, Russa; Bjerva, Johannes; Hose, Katja. 2024. Knowledge graphs, large language models, and hallucinations: An NLP perspective. *arXiv preprint* arXiv:2411.14258. Available online: https://arxiv.org/abs/2411.14258 (accessed December 19, 2025).

Lee, Lenka; Mácha, Jakub. 2024. Inverted ekphrasis and hallucinating stochastic parrots: Deleuzean insights into AI and art in daily life. *Itinera* 28. https://doi.org/10.54103/2039-9251/27840.

Lewandowsky, Stephan; Ecker, Ullrich K. H.; Seifert, Colleen M.; Schwarz, Norbert; Cook, John. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest* 13 (3), 106–131. https://doi.org/10.1177/1529100612451018. Available online: https://pubmed.ncbi.nlm.nih.gov/22922134/ (accessed December 25, 2025).

Li, Kenneth; Hopkins, Aspen K.; Bau, David; Viégas, Fernanda; Pfister, Hanspeter; Wattenberg, Martin. 2022. *Emergent world representations: Exploring a sequence model trained on a synthetic task*. *arXiv preprint* arXiv:2210.13382. Available online: https://arxiv.org/abs/2210.13382 (accessed December 25, 2025).

Lin, Zhen; Fu, Yao; Zhang, Ben; Zhang, Tianyi; Chen, Danqi. 2024. FLAME: Factuality-aware alignment for large language models. *Advances in Neural Information Processing Systems* 37. Available online: https://www.proceedings.com/079017-3671.html (accessed December 19, 2025).

Meng, Kevin; Bau, David; Andonian, Alex; Belinkov, Yonatan. 2023. Locating and editing factual associations in GPT. In *Proceedings of ICLR 2023*. Available online: https://arxiv.org/abs/2202.05262 (accessed December 19, 2025).

Michel, Jean-Baptiste; Shen, Yuan Kui; Aiden, Aviva Presser; et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331, 176–182. https://doi.org/10.1126/science.1199644.

Min, Sewon; Krishna, Kalpesh; Lyu, Xinxi; et al. 2023. FactScore: Fine-grained atomic evaluation of factual precision in long-form generation. *arXiv preprint* arXiv:2305.14251. Available online: https://arxiv.org/abs/2305.14251 (accessed December 19, 2025).

Mitchell, Eric; Lin, Charles; Bosselut, Antoine; Manning, Christopher D.; Finn, Chelsea. 2022. Memory-based model editing at scale. *arXiv preprint* arXiv:2206.06520. Available online: https://arxiv.org/abs/2206.06520 (accessed December 25, 2025).

Mollo, Dimitri; Millière, Raphaël. 2025. The vector grounding problem. *arXiv preprint arXiv*:2304.01481v3. Available online: https://arxiv.org/abs/2304.01481 (accessed December 19, 2025). Forthcoming in *Philosophy and the Mind Sciences*.

Mulligan, Kevin; Simons, Peter; Smith, Barry. 1984. Truth-makers. *Philosophy and Phenomenological Research* 44, 287–321.

Olah, Chris; Cammarata, Nick; Schubert, Ludwig; Goh, Gabriel; Petrov, Michael; Carter, Shan. 2020. *Zoom in: An introduction to circuits*. *Distill* 5 (3). Available online: https://distill.pub/2020/circuits/zoom-in/ (accessed December 25, 2025).

Putnam, Hilary. 1981. *Reason, truth and history*. Cambridge: Cambridge University Press.

Quine, W. V. O. 1951. Two dogmas of empiricism. *Philosophical Review* 60, 20–43.

Russell, Bertrand. 1912. *The problems of philosophy*. London: Williams and Norgate.

 doi:10.20944/preprints202601.1447.v1

Russell, Bertrand. 1918/2010. *The philosophy of logical atomism*. London: Routledge.

Russell, Stuart J.; Norvig, Peter. 2010. *Artificial intelligence: A modern approach*, 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Ryle, Gilbert. 1949. *The concept of mind*. London: Hutchinson.

Sansford, Hannah; Richardson, Nicholas; Petric Maretic, Hermina; Nait Saada, Juba. 2024. *GraphEval: A knowledge-graph based LLM hallucination evaluation framework*. *arXiv preprint* arXiv:2407.10793. Available online: https://arxiv.org/abs/2407.10793 (accessed December 25, 2025).

Saussure, Ferdinand de. 1983. *Course in General Linguistics*. Translated by Roy Harris. London: Duckworth.

Šekrst, Kristina. Forthcoming. Do large language models hallucinate electric fata morganas? *Journal of Consciousness Studies*.

Sharkey, Lee; Chughtai, Bilal; Batson, Joshua; et al. 2025. Open problems in mechanistic interpretability. *arXiv preprint arXiv*:2501.16496. Available online: https://arxiv.org/abs/2501.16496 (accessed December 25, 2025).

Templeton, Adly; et al. 2024. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits. Available online: https://transformer-circuits.pub/2024/scaling-monosemanticity/ (accessed December 25, 2025).

Trinh, Tue. 2024. Logicality and the picture theory of language: Propositions as pictures in Wittgenstein's *Tractatus*. *Synthese* 203, 127. https://doi.org/10.1007/s11229-024-04549-4.

Tuquero, Loreben. 2025. Musk's AI-powered Grokipedia: A Wikipedia spin-off with less care to sourcing, accuracy. *PolitiFact*, November 12, 2025. Available online: https://www.politifact.com/article/2025/nov/12/Grokipedia-Wikipedia-AI-citations/ (accessed December 26, 2025).

Turpin, Miles; Michael, Julian; Perez, Ethan; Bowman, Samuel R. 2023. *Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting*. In *Advances in Neural Information Processing Systems* 36. https://doi.org/10.48550/arXiv.2305.04388. Available online: https://arxiv.org/abs/2305.04388 (accessed December 25, 2025).

Verheggen, Claudine; Myers, Robert H. 2025. The status and the scope of the principle of charity. *Topoi* 44, 1215–1226. https://doi.org/10.1007/s11245-025-10275-4.

Weatherby, Leif. 2025. *Language machines: Cultural AI and the end of remainder humanism*. Minneapolis: University of Minnesota Press.

Wei, Haoran; Sun, Yaofeng; Li, Yukun. 2025. DeepSeek-OCR: Contexts optical compression. *arXiv preprint arXiv*:2510.18234. Available online: https://arxiv.org/abs/2510.18234 (accessed December 25, 2025).

Wittgenstein, Ludwig. 1922. *Tractatus logico-philosophicus*. Trans. by C. K. Ogden. London: Kegan Paul.

Wittgenstein, Ludwig. 1929. Some remarks on logical form. *Proceedings of the Aristotelian Society, Supplementary Volume* 9, 162–171. https://doi.org/10.1093/aristoteliansupp/9.1.162. Available online: https://www.jstor.org/stable/4106481 (accessed December 19, 2025).

Wittgenstein, Ludwig. 1980. *Wittgenstein's lectures, Cambridge, 1930–1932: From the notes of John King and Desmond Lee*. Edited by D. Lee. Oxford and Chicago: Basil Blackwell and University of Chicago Press.

Xu, Ziwei; Jain, Sanjay; Kankanhalli, Mohan. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint* arXiv:2401.11817. Available online: https://arxiv.org/abs/2401.11817 (accessed December 19, 2025).

Yao, Liang; Mao, Chengsheng; Luo, Yuan. 2019. *KG-BERT: BERT for knowledge graph completion*. *arXiv preprint* arXiv:1909.03193. Available online: https://arxiv.org/abs/1909.03193 (accessed December 25, 2025).

Zhang, Yuji; Li, Sha; Liu, Jiateng; Yu, Pengfei; Fung, Yi R.; Li, Jing; Li, Manling; Ji, Heng. 2024. *Knowledge overshadowing causes amalgamated hallucination in large language models*. *arXiv preprint* arXiv:2407.08039. Available online: https://arxiv.org/abs/2407.08039 (accessed December 25, 2025).

Zhang, Yuji; Li, Sha; Qian, Cheng; Liu, Jiateng; Yu, Pengfei; Han, Chi; Fung, Yi R.; McKeown, Kathleen; Zhai, Chengxiang; Li, Manling; Ji, Heng. 2025. *The Law of Knowledge Overshadowing: Towards Understanding, Predicting, and Preventing LLM Hallucination*. *arXiv preprint* arXiv:2502.16143. Available online: https://arxiv.org/abs/2502.16143 (accessed December 25, 2025).

 doi:10.20944/preprints202601.1447.v1

Zheng, Liwen; Li, Chaozhuo; Liu, Zheng; Huang, Feiran; Jia, Haoran; Ye, Zaisheng; Zhang, Xi. 2025. Fact in fragments: Deconstructing complex claims via LLM-based atomic fact extraction and verification. arXiv preprint arXiv:2506.07446. Available online: https://arxiv.org/abs/2506.07446 (accessed December 25, 2025).

Zou, Andy; Phan, Long; Chen, Sarah; et al. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint* arXiv:2310.01405. Available online: https://arxiv.org/abs/2310.01405 (accessed December 19, 2025).