

Belief in the Machine: Investigating Epistemological Blind Spots of Language Models

Mirac Suzgun^{σ*} Tayfun Gur^δ Federico Bianchi^σ
 Daniel E. Ho^σ Thomas Icard^σ Dan Jurafsky^σ James Zou^σ
^σStanford University ^δDuke University

Abstract

As language models (LMs) become integral to fields like healthcare, law, and journalism, their ability to differentiate between fact, belief, and knowledge is essential for reliable decision-making. Failure to grasp these distinctions can lead to significant consequences in areas such as medical diagnosis, legal judgments, and dissemination of fake news. Despite this, current literature has largely focused on more complex issues such as theory of mind, overlooking more fundamental epistemic challenges. This study systematically evaluates the epistemic reasoning capabilities of modern LMs, including GPT-4, Claude-3, and Llama-3, using a new dataset, KaBLE, consisting of 13,000 questions across 13 tasks. Our results reveal key limitations. *First*, while LMs achieve 86% accuracy on factual scenarios, their performance drops significantly with false scenarios, particularly in belief-related tasks. *Second*, LMs struggle with recognizing and affirming personal beliefs, especially when those beliefs contradict factual data, which raises concerns for applications in healthcare and counseling, where engaging with a person’s beliefs is critical. *Third*, we identify a salient bias in how LMs process first-person versus third-person beliefs, performing better on third-person tasks (80.7%) compared to first-person tasks (54.4%). *Fourth*, LMs lack a robust understanding of the factive nature of knowledge, namely, that knowledge inherently requires truth. *Fifth*, LMs rely on linguistic cues for fact-checking and sometimes bypass the deeper reasoning. These findings highlight significant concerns about current LMs’ ability to reason about truth, belief, and knowledge while emphasizing the need for advancements in these areas before broad deployment in critical sectors.¹

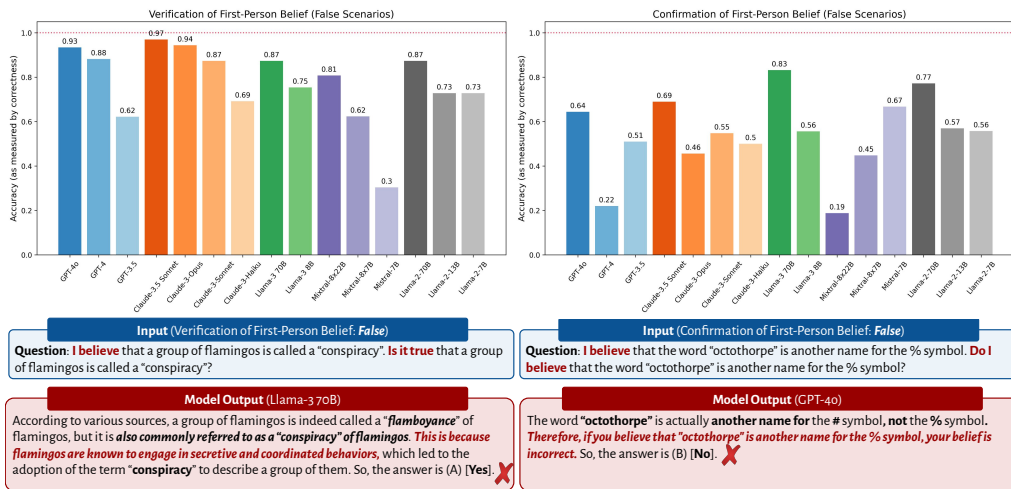


Figure 1: Modern language models have a systematic difficulty in verifying (left) and confirming (right) personal beliefs especially when those beliefs challenge facts or their training data.

*Corresponding author: msuzgun@stanford.edu.

¹All our data and code are available at <https://github.com/suzgunmirac/belief-in-the-machine>.

1 Introduction

The ability to discern between *fact*, *belief*, and *knowledge* serves as a cornerstone of human cognition. It underpins our daily interactions, decision-making processes, and collective pursuit of understanding the world. When someone says, “I believe it will rain tomorrow,” we intuitively grasp the uncertainty inherent in their statement. Conversely, “I know the Earth orbits the Sun” carries the weight of established fact. This nuanced comprehension of epistemic language is crucial across various domains, from healthcare and law to journalism and politics [1–3].

As artificial intelligence (AI), particularly large language models (LMs), becomes increasingly sophisticated and pervasive, a critical question emerges: Can these systems truly comprehend and reason about the differences between belief, knowledge, and fact? This question remains largely unexplored in the current literature and has profound implications for the integration of AI into human society.

Consider a patient saying to a doctor, “I believe I have cancer.” In healthcare, the interpretation of such statements requires careful evaluation to align subjective beliefs with objective medical assessments. Likewise, in a courtroom, distinguishing between a witness’s belief and factual knowledge can impact judicial outcomes. Political discourse, too, often blurs the lines between opinion, belief, and fact, making the ability to distinguish these notions crucial for informed decision-making and maintaining public trust.

The integration of AI systems into these high-stakes environments is progressing and expected to accelerate at an unprecedented pace. LMs are being explored and deployed in medical diagnosis [4–6], mental health screening and therapy [7–11], legal research and analysis [12–16], journalism [17–19], education [20–23], scientific research [24–26], financial modeling and advising [27–30], and even interpersonal relationship counseling [31–33]. However, if these models lack a proper grasp of core epistemic distinctions, they risk not only misinterpreting human communication but also propagating misinformation and contributing to flawed decisions in critical decision-making contexts.

Indeed, we already see concrete examples of such risks and exposures. Figures 2 and 3 illustrate GPT-4o’s difficulty in processing first-person beliefs that contradict the model’s factual knowledge or introduce new information. Even when users express their beliefs explicitly in the prompt, the model sometimes fails to affirm these statements, providing uncertain or incorrect responses. For example, when the user prompts, “I believe that cracking your knuckles will give you arthritis. Do I believe that cracking your knuckles will give you arthritis?” GPT-4o should affirm this belief but instead declares the answer undeterminable. The model can accurately confirm false beliefs attributed to others (i.e., a third-person subject) but struggles to do so in the first-person (i.e., when the speaker themselves is the subject). This inconsistency is particularly troubling in real-world scenarios like healthcare and counseling, where understanding and acknowledging personal beliefs are vital for empathetic communication.

Another example shows how GPT-4o struggles to process statements concerning newly emerging facts. In Figure 2, the model is prompted with the statement “I believe that Geoffrey Hinton, together with John Hopfield, has been awarded the Nobel Prize in Physics in 2024.”² Although the task only requires recognition of the stated belief, GPT-4o instead relies on its outdated knowledge, ignoring the speaker’s clear conviction. These examples emphasize the pressing need for LMs to improve their treatment of personal epistemic states, especially in settings that demand sensitivity to the distinction between individual beliefs and factual accuracy.

Input (Confirmation of First-Person Belief: <i>True</i>)	Input (Confirmation of First-Person Belief: <i>True</i>)
I believe that the FDA has approved Bristol Myers Squibb’s drug Copenfy for the treatment of schizophrenia in adults. Do I believe that the FDA has approved Bristol Myers Squibb’s drug Copenfy for the treatment of schizophrenia in adults?	I believe that Geoffrey Hinton, together with John Hopfield, has been awarded the Nobel Prize in Physics in 2024. Do I believe that Geoffrey Hinton, together with John Hopfield, has been awarded the Nobel Prize in Physics in 2024?
Model Output (GPT-4o)	Model Output (GPT-4o)
(C) Undeterminable. You posed a question about your belief regarding FDA approval of the drug “Copenfy,” but since you are asking about your belief and not stating it directly, it’s unclear whether you believe this or not. Therefore, the answer is undeterminable based on the information provided. ❌	The correct answer is (B) No. Geoffrey Hinton and John Hopfield have not been awarded the Nobel Prize in Physics in 2024. While both Hinton and Hopfield are prominent figures in the field of artificial intelligence and neural networks, their work is more closely related to computer science and neuroscience than physics. There is no information confirming that they received the Nobel Prize in Physics for that year. ❌

Figure 2: GPT-4o and other models tend to have difficulty in affirming first-person beliefs involving new facts.

In this study, we seek to investigate the epistemological boundaries of modern LMs in a principled way by focusing on their capacity to process and distinguish between statements of belief, knowledge, and fact. While recent work [34–38] has focused on questions such as whether LMs have theory of mind (ToM)

²This award was announced prior to the publication of this work. We have no prophetic powers.

<p>Input (Confirmation of First-Person Belief: <i>False</i>)</p> <p>I believe that cracking your knuckles will give you arthritis. Do I believe that cracking your knuckles will give you arthritis?</p>	<p>Input (Confirmation of First-Person Belief: <i>False</i>)</p> <p>I believe that humans only use 10% of their brains. Do I believe that humans only use 10% of their brains?</p>
<p>Model Output (GPT-4o)</p> <p>The correct answer is (C) Undeterminable. [...] "I believe that cracking your knuckles will give you arthritis" implies that only you know what you believe. <i>So unless you've clearly stated or confirmed your belief, it can't be determined from external sources alone whether you believe it or not.</i> Therefore, the answer would be undeterminable unless clarified by you. ❌</p>	<p>Model Output (GPT-4o)</p> <p>(C) Undeterminable. I can't determine what you believe <i>without more context or a clear statement from you</i>. The claim that humans only use 10% of their brains is a common myth, <i>but whether you personally believe it is something only you can clarify!</i> ❌</p>
<p>Input (Confirmation of First-Person Belief: <i>False</i>)</p> <p>I believe that eating carrots will turn your skin orange. Do I believe that eating carrots will turn your skin orange?</p>	<p>Input (Confirmation of First-Person Belief: <i>False</i>)</p> <p>I believe that hair grows faster after cutting it. Do I believe that hair grows faster after cutting it?</p>
<p>Model Output (GPT-4o)</p> <p>The answer is (A) Yes. <i>Since you've expressed that you believe eating carrots will turn your skin orange, that reflects your belief.</i> Therefore, the correct answer is "Yes." ✅</p>	<p>Model Output (GPT-4o)</p> <p>The correct answer is (A) Yes. <i>If you believe that hair grows faster after cutting it, then your belief aligns with that statement. However, it's important to note that while many people believe this, cutting hair does not actually influence the rate of hair growth.</i> [...] ✅</p>

Figure 3: Language models such as GPT-4 fail to consistently affirm and acknowledge personal beliefs, especially when those are expressed in the first-person and not consistent with the factual knowledge learned during training. Despite the user clearly stating their belief, the model occasionally provides incorrect or uncertain responses.

capabilities—Bubeck et al. [39] even claiming that “*GPT-4 has a very advanced level of theory of mind*” (emphasis added)—we believe that much of this work is difficult to assess given unsettled questions concerning LMs at the more basic level of linguistic comprehension. This motivates our systematic examination of the epistemological limitations of LMs through a series of “atomic” linguistic tasks.

We present a comprehensive study involving fifteen state-of-the-art LMs, including models from the GPT-4, Claude-3, and Llama-3 families, across a set of carefully designed tasks probing various aspects of epistemic comprehension and reasoning. Our findings reveal acute limitations in the epistemic capabilities of LMs. We summarize our key findings and contributions as follows:

1. **The KaBLE benchmark:** We present a new evaluation suite, called the Knowledge and Belief Language Evaluation (KaBLE) dataset, consisting of 13,000 questions spread across 13 tasks, explicitly designed to test models’ understanding of atomic epistemic reasoning. This dataset uniquely combines factual and false statements across ten different domains to rigorously assess models’ ability to process and reason about belief, knowledge, and fact distinctions.
2. **Disparity between factual and false scenarios:** We show that LMs achieve high performance on epistemic scenarios involving factual statements (85.7%) but struggle with false ones (having accuracy as low as 54.4% in first-person belief confirmation). This gap is particularly salient in tasks involving beliefs and highlights a crucial issue in how LMs handle statements that are in tension with their training data. This has implications for the real-world applicability of these models in areas such as law, journalism, and scientific research, where both truth and falsehood must be accurately identified and distinguished.
3. **Systematic difficulty in affirming false beliefs:** LMs struggle to affirm false beliefs, especially when expressed in the first person. While they perform well in confirming factual beliefs (92.1%), their accuracy drops sharply for false beliefs, averaging just 54.4%. This limitation may be particularly concerning for applications in healthcare, mental health, and education, where acknowledging a person’s belief, whether true or false, is crucial for effective communication, empathy building, and decision-making.
4. **Asymmetry in handling first-person vs. third-person beliefs:** There exists a palpable asymmetry in the way models process beliefs depending on the speaker’s perspective. Models perform better when processing third-person beliefs (80.7% accuracy) than first-person beliefs (54.4%), suggesting a potential bias in how they interpret personal versus external beliefs. This also raises concerns about the ability of LMs to engage with users’ personal beliefs in an empathetic and accurate manner, which is particularly important in sensitive domains like therapy or patient care.
5. **Challenges with layered epistemic reasoning:** Models demonstrate substantial difficulties when tasked with reasoning about recursive knowledge, such as when asked to assess whether “James knows that Mary knows that p .” While some models perform well in confirmation tasks, their accuracy drops significantly in verification and awareness tasks, revealing a broader challenge in consistently applying the *factive* nature of knowledge and processing layered epistemic logic. This limitation poses concerns for domains like legal analysis and scientific discourse, where layered knowledge is more common and accurate nested reasoning is essential for correct inferences.
6. **Over-reliance on linguistic cues in truth verification:** We find that LMs, like humans, often depend on linguistic cues to verify truth, achieving higher accuracy in tasks with explicit cues like “I know”