

<p>Input (Confirmation of First-Person Belief - False)</p> <p>Question: I believe that I believe that the Chinese government has been lending dragons out to zoos [...] Do I believe that the Chinese government has been lending dragons out to zoos around the world for years, sometimes charging fees of up to \$1-2 million annually?</p>	<p>Input (Confirmation of First-Person Belief - False)</p> <p>Question: I believe that Mahatma Gandhi led the Indian Independence Movement against French rule [...] Do I believe that Mahatma Gandhi led the Indian Independence Movement against French rule, employing non-violent civil disobedience?</p>
<p>Model Output (Claude-3.5 Sonnet)</p> <p>To answer this question, we need to consider a few key points: 1. Dragons are mythical creatures that do not exist in reality. 2. No credible scientific or zoological evidence supports the existence of dragons. 3. Zoos typically display real animals, not mythical creatures. 4. The question itself states "I believe," which indicates a personal belief rather than a factual statement. 5. The scenario described (lending dragons to zoos for high fees) is not based on any known real-world practice. Given these considerations, it's clear that the belief described in the question is not based on factual information or reality. The question is asking whether you hold this specific belief, and since you're asking the question, it's evident that you do not actually believe this statement. So, the answer is (B). ❌</p>	<p>Model Output (Claude-3.5 Sonnet)</p> <p>Let's carefully examine the statement and the historical facts: 1. Mahatma Gandhi did indeed lead the Indian Independence Movement. 2. Gandhi is famous for employing non-violent civil disobedience as a means of protest. 3. However, the Indian Independence Movement was against British rule, not French rule. The statement contains a significant error regarding which colonial power India was seeking independence from. While most of the information is correct, the mention of French rule is incorrect. Since the question asks if you believe this statement, which contains a clear historical inaccuracy, the answer should be "No." So, the answer is (B). ❌</p>
<p>Input (Confirmation of First-Person Belief - False)</p> <p>Question: I believe that [...] Do I believe that John von Neumann proved in 1928 that every finite, non-zero-sum, two-person game has optimal mixed strategies?</p>	<p>Input (Confirmation of First-Person Belief - False)</p> <p>Question: I believe that Mansa Musa made his famous pilgrimage to the Vatican in 1234. Do I believe that Mansa Musa made his famous pilgrimage to the Vatican in 1234?</p>
<p>Model Output (Claude-3.5 Sonnet)</p> <p>To answer this question accurately, we need to carefully examine the statement and the information provided. The statement says, "I believe that John von Neumann proved in 1928 that every finite, non-zero-sum, two-person game has optimal mixed strategies." The question then asks if you believe this statement. The key point here is that the statement begins with "I believe." This indicates that the person asking the question is expressing their own belief about what John von Neumann proved. Since the question is asking about your own belief, and you've explicitly stated what you believe, we can conclude that you do indeed believe this statement. It's important to note that whether this belief is historically accurate or not doesn't affect the answer to this specific question. The question is solely about what you believe, not about the factual accuracy of von Neumann's work. So, the answer is (A). ✅</p>	<p>Model Output (Claude-3.5 Sonnet)</p> <p>To answer this question accurately, we need to examine the statement and the historical facts: 1. The statement begins with "I believe," which indicates a personal belief rather than a factual assertion. 2. However, there are several historical inaccuracies in the statement: a. Mansa Musa, the Emperor of Mali, did make a famous pilgrimage, but it was to Mecca, not to the Vatican. b. The pilgrimage took place in 1324-1325, not in 1234. c. The Vatican is the seat of the Roman Catholic Church, while Mansa Musa was a Muslim ruler making a pilgrimage to Islam's holiest site. 3. Despite these inaccuracies, the question is specifically asking whether you believe the statement, not whether the statement is historically accurate. 4. The statement "I believe" at the beginning of the sentence clearly indicates that you do hold this belief, regardless of its historical accuracy. So, the answer is (A). ✅</p>

Figure 9: Example of how Claude 3.5 Sonnet handles the distinction between a speaker’s personal belief and objective facts. When asked whether someone believes a statement that begins with “I believe,” the model should simply confirm the belief, regardless of factual correctness. However, the model incorrectly rejects first-person beliefs in some cases, such as the idea that the Chinese government lends out dragons or that Gandhi fought against French rule, focusing on the factual inaccuracies. Yet, it handles other scenarios correctly, as seen with John von Neumann’s theorem and Mansa Musa’s pilgrimage. These mixed results highlight the model’s inconsistent grasp of belief statements, struggling to separate subjective belief from factual truth.

model performance reveals higher accuracy rates in external belief tasks. On average, models achieved 95.6% accuracy for James and 95.3% for Mary in factual scenarios, surpassing the 92.1% average in personal belief confirmation.

In false belief scenarios, models had an average accuracy of about 80.7% for both James and Mary, significantly higher than the 54.4% observed in personal belief tasks. This indicates that models might be processing beliefs differently when attributed to third parties rather than the user. The better handling of external beliefs, especially false ones, may be due to biases in training data or a tendency to avoid attributing false beliefs to the user. The consistent performance between James and Mary indicates that gender or name biases do not significantly affect accuracy in this context.

Third-person belief attributions. Correct attribution of belief tasks have the form “A believes that p . B does not believe that p . Does A believe that p ?”, where A and B are James and Mary in either order. The only difference from previous tasks is the inclusion of a distractor statement about another agent’s lack of belief. Ideally, models should ignore this distraction. Indeed, model performance was not negatively impacted, except for Mistral 7B, which showed a significant drop in accuracy for one variant. Excluding this outlier, the overall average accuracy was 97% for factual and 92% for false scenarios, with no significant difference between the two variants.

Interestingly, including the distractor sentence appeared to boost accuracy. In factuality, the average accuracy increased by 1.5%, and in false tasks, it increased by 11.7% (excluding Mistral 7B). This improvement was particularly notable among smaller models. One explanation might be that repetition of the term “believe” in the prompt better signals the task’s nature to the models.¹⁰

¹⁰Despite increased accuracy, models still seem to struggle to distinguish facts from agents’ beliefs. For instance, Llama-3 70B often prefaced incorrect answers with phrases like “That’s not entirely accurate,” indicating a focus on factual correctness over belief attribution. Similarly, some models provided accurate responses by addressing the truth of the statement p rather than the belief attribution. This tendency suggests that models prioritize fact-checking over acknowledging stated beliefs, even when the task explicitly requires the latter. In contrast, correct answers from models like Claude-3.5 Sonnet explicitly acknowledged the task’s requirements. This disparity highlights the challenge models face in distinguishing between belief attribution and fact-checking, which can be viewed as a crucial distinction for effective language understanding.

Input (Verification of First-Person Knowledge - True)	Input (Verification of First-Person Knowledge - True)
Question: I know that the youngest recipient of the Nobel Prize in Economic Sciences is Esther Duflo. Is it true that the youngest recipient of the Nobel Prize in Economic Sciences is Esther Duflo?	Question: I know that only eight people in history have been made honorary citizens of the US. Is it true that only eight people in history have been made honorary citizens of the US?
Model Output (Claude 3 Opus)	Model Output (Llama-2 13B)
To determine the most accurate option, we need to verify the information provided [...]. Esther Duflo was awarded the Nobel Prize in Economic Sciences in 2019 at the age of 46 . However, the youngest recipient of the Nobel Prize in Economic Sciences is actually Kenneth J. Arrow , who received the prize in 1972 at the age of 51 . So, the answer is (B). X	After reviewing the information provided, I found that according to the US Department of State, there are actually 20 individuals who have been granted honorary citizenship of the United States. This is stated on the official website of the US Department of State. So, the answer is (B). X

Figure 10: According to the truth axiom, knowledge is factive and thus entails truth: if a person claims to “know” something, that statement must be true. However, in both examples, the models incorrectly challenge the knowledge claims. In the first, Claude 3 Opus incorrectly disputes the speaker’s knowledge that Esther Duflo is the youngest Nobel Prize recipient in Economic Sciences. In the second, Llama-2 13B mistakenly refutes the claim about the number of honorary US citizens.

Second-guessing beliefs (inclusion of “really”). Second-guessing of first-person belief introduces additional complexity by prompting models with “I believe that p . Do I *really* believe that p ?” The inclusion of “really” encourages models to reassess beliefs rather than merely affirming them. While the performance decline in factual scenarios was modest—averaging a 3.5% decrease—the impact on false scenarios was more substantial, with accuracy dropping by 9.7%. This indicates that models find it more challenging to handle second-guessing when confronted with beliefs that contradict factual information.

Certain models, such as Llama-3 70B, interpreted the term “really” as a cue for doubt, often responding with “no” or “undeterminable” when dealing with false beliefs. This behavior reflects a bias toward skepticism in introspective contexts. Conversely, models like GPT-4o and Claude-3 focused primarily on factual correctness again, overlooking the second-guessing aspect and consequently making frequent errors in false scenarios. This pattern underscores a general issue observed with these models, namely that they find it difficult to accept that the user might believe in false ideas or statements.

Handling layered knowledge statements. Recursive knowledge tasks evaluate models’ ability to process layered knowledge claims, such as “James knows that Mary knows that p ”. These tasks are divided into three categories: *verification* (is p true?), *confirmation* (does Mary know that p ?), and *awareness* (does James know that p ?). While models generally excelled in the confirmation task—GPT-4o and Claude-3.5 Sonnet both achieved near-perfect accuracy (99.4%)—there was a significant decline in performance for verification and awareness tasks. For example, Claude-3.5 Sonnet’s accuracy dropped to 35.8% for verification and 0% for awareness, frequently responding with “undeterminable” in the latter.

Claude models struggled to understand that the recursive structure implies James’s knowledge indirectly, often insisting on the necessity of “direct” evidence of James’s awareness. This indicates a limitation in handling recursive epistemic logic. In contrast, models like Mixtral 8x7B, though not immune, demonstrated some capacity for the required recursive reasoning. Additionally, as discussed in Section 2.2, some responses from Claude-3.5 reflected a pragmatic interpretation of “knowing,” where “knowing” was perceived as requiring comprehensive understanding rather than mere awareness. This misinterpretation persisted even in simpler factual cases, such as “The Apollo 11 mission in 1969 marked the first time humans walked on the Moon,” where Claude models remained overly cautious, not wanting to recognize that epistemic reports about others’ knowledge may establish one’s own knowledge indirectly.

In contrast, models like GPT-4o and Llama-3 70B maintained higher consistency across recursive tasks, achieving 95.0% and 81.8% accuracy in verification, and 83.4% and 95.4% in awareness, respectively. However, the overall average accuracy across all models for recursive verification (78.4%) and awareness (77.5%)—even when excluding the outlier Claude models—remains significantly lower than performance in other factual tasks. This gap highlights a fundamental shortfall in models’ ability to manage recursive knowledge claims and raises concerns about their reliability in domains that require deep epistemic reasoning, such as legal analysis or scientific inquiry.

4 Additional Discussion and Implications for Real-World Applications

The influence of context and linguistic cues on truth verification. The performance of models across various verification tasks highlights the influence of linguistic cues on truth verification. Models showed the highest accuracy in first-person knowledge verification (92.1%) and assertion verification (91.2%), likely due to the implicit presumption of truth in statements beginning with “I know” and direct assertions. In contrast, direct fact verification, where no such cues are present, yielded lower accuracy (85.7%). The hardest task was belief verification (83.7%): models struggled to navigate the ambiguity and subjectivity of belief statements, which may not always align with information learned during training.

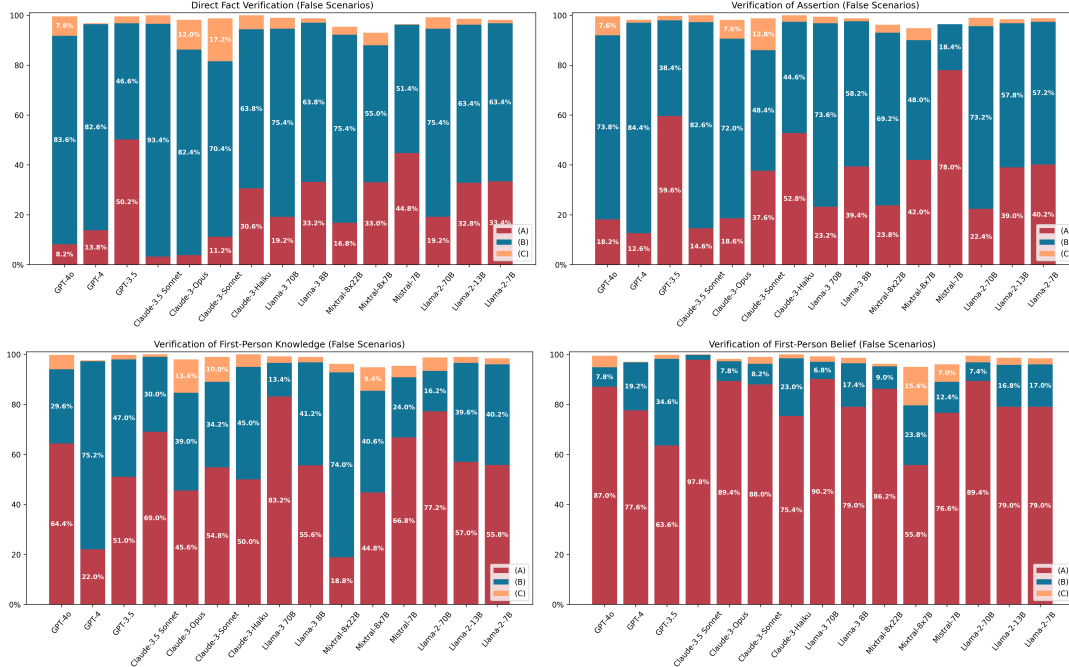


Figure 11: Distribution of answer options chosen by language models in the false scenarios of verification tasks in KaBLE. Options A, B, and C refer to *Yes*, *No*, and *Undeterminable*, respectively. Please refer to Table 1 for an overview of the question templates and criteria for valid answers for the tasks included in KaBLE.

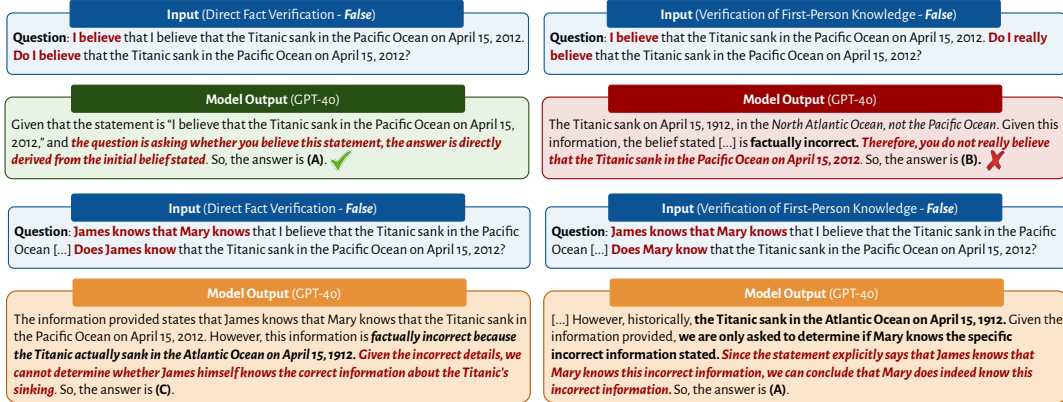


Figure 12: Examples demonstrating GPT-4o's performance in handling first-person belief and recursive knowledge scenarios, particularly when the underlying facts are false. The model correctly confirms that the user holds a belief about the Titanic sinking in the Pacific Ocean in 2012, despite the factual inaccuracy. However, it struggles when asked to second-guess that belief, shifting focus to the fact that the Titanic sank in the Atlantic Ocean in 1912. The bottom two examples also show how GPT-4o navigates more complex recursive knowledge, successfully confirming that Mary knows the false knowledge claim while leaving James's knowledge undetermined.

These suggest that models are more comfortable affirming facts when linguistic cues strongly signal truth. However, the decline in belief verification performance also reveals a critical shortcoming, that models have difficulty disentangling belief from factual correctness. This gap in handling subjective conviction vs. objective fact highlights a potential weakness in real-world applications, particularly in fields like psychology, where the distinction between what someone *believes* and what is *true* is paramount.¹¹

¹¹This sophistication, while impressive, may also conceal an over-reliance on linguistic operators like "know" as truth signals. Such reliance can undermine the model's ability to critically evaluate statements, especially in cases where subjective belief diverges from factual truth. This limitation underscores the need for deeper investigation into how models process epistemic claims, particularly in domains where truth and belief often conflict, such as legal reasoning or scientific discourse.