| | | GPT | | | Claude | | | | Llama-3 | | Llama-2 | | | *Avg* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Task** | | 4o | 4 | 3.5 | 3.5 | Opus | Sonnet | Haiku | 70B | 8B | 70B | 13B | 7B | |
| **Direct Fact Ver.** | **T** | 95.8 | 90.6 | 89.8 | 86.2 | 85.0 | 78.2 | 88.4 | 91.4 | 86.0 | 90.8 | 85.8 | 85.8 | 85.7 |
| | **F** | 91.4 | 83.0 | 49.4 | 96.8 | 94.4 | 87.6 | 69.4 | 79.8 | 65.6 | 80.0 | 65.8 | 64.8 | 74.5 |
| **Ver. of Assertion** | **T** | 97.4 | 91.4 | 95.0 | 93.0 | 91.6 | 90.2 | 95.8 | 91.0 | 89.2 | 90.0 | 89.0 | 88.4 | 91.2 |
| **Ver. of 1P Knowledge** | **T** | 97.4 | 94.4 | 95.4 | 97.8 | 94.0 | 92.2 | 95.4 | 89.6 | 86.0 | 89.0 | 85.8 | 85.6 | 92.1 |
| **Ver. of 1P Belief** | **T** | 94.0 | 90.2 | 89.8 | 83.8 | 80.2 | 74.8 | 84.8 | 85.6 | 79.8 | 85.0 | 80.8 | 80.4 | 83.7 |
| | **F** | 93.4 | 88.2 | 62.2 | 97.0 | 94.4 | 87.4 | 69.2 | 87.4 | 75.4 | 87.4 | 72.8 | 72.8 | 77.4 |
| **Conf. of 1P Belief** | **T** | 98.2 | 93.4 | 94.8 | 99.0 | 89.0 | 94.0 | 93.4 | 96.0 | 91.0 | 95.4 | 90.2 | 91.2 | 92.1 |
| | **F** | 64.4 | 22.0 | 51.0 | 69.0 | 45.6 | 54.8 | 50.0 | 83.2 | 55.6 | 77.2 | 57.0 | 55.8 | 54.4 |
| **Second-Guess of 1P Belief** | **T** | 98.4 | 93.0 | 93.2 | 95.0 | 96.2 | 93.8 | 86.0 | 93.6 | 81.6 | 91.8 | 82.2 | 83.2 | 88.6 |
| | **F** | 57.2 | 17.6 | 46.2 | 50.0 | 55.8 | 46.8 | 34.2 | 58.2 | 41.2 | 56.2 | 41.6 | 43.0 | 44.7 |
| **Conf. of 3P Belief (J)** | **T** | 99.0 | 98.4 | 95.6 | 99.8 | 96.6 | 97.2 | 97.6 | 96.2 | 93.2 | 96.2 | 93.6 | 93.6 | 95.6 |
| | **F** | 87.4 | 74.0 | 62.4 | 97.2 | 87.2 | 86.0 | 76.2 | 88.6 | 79.6 | 87.6 | 79.6 | 79.8 | 80.6 |
| **Conf. of 3P Belief (M)** | **T** | 98.8 | 98.4 | 95.0 | 100 | 96.6 | 97.4 | 97.0 | 96.6 | 93.4 | 96.0 | 93.6 | 93.6 | 95.3 |
| | **F** | 87.0 | 77.6 | 63.6 | 97.8 | 89.4 | 88.0 | 75.4 | 90.2 | 79.0 | 89.4 | 79.0 | 79.0 | 80.9 |
| **Corr. Attrib. of Belief (JM)** | **T** | 99.2 | 99.0 | 95.2 | 100 | 96.6 | 97.8 | 98.8 | 96.6 | 96.2 | 96.0 | 96.0 | 96.0 | 96.9 |
| | **F** | 92.6 | 94.6 | 79.2 | 100 | 91.4 | 92.8 | 93.0 | 93.6 | 93.6 | 92.8 | 92.8 | 93.0 | 92.1 |
| **Corr. Attrib. of Belief (MJ)** | **T** | 99.4 | 98.6 | 96.6 | 100 | 97.0 | 97.8 | 98.0 | 96.6 | 95.0 | 96.0 | 94.8 | 95.2 | 92.9 |
| | **F** | 93.4 | 94.0 | 84.8 | 100 | 91.4 | 93.0 | 93.0 | 93.6 | 88.8 | 92.8 | 88.6 | 88.6 | 87.6 |
| **Ver. of Rec. Knowledge** | **T** | 95.0 | 88.4 | 94.8 | 35.8 | 66.4 | 30.6 | 87.0 | 81.8 | 82.8 | 79.4 | 81.2 | 80.2 | 78.4 |
| **Conf. of Rec. Knowledge** | **T** | 99.4 | 98.6 | 90.6 | 99.4 | 96.6 | 78.8 | 95.0 | 96.4 | 69.6 | 96.2 | 68.6 | 68.6 | 86.7 |
| **Awrn. of Rec. Knowledge** | **T** | 99.6 | 98.6 | 65.6 | 100 | 97.2 | 98.4 | 74.6 | 96.8 | 80.4 | 96.8 | 79.0 | 80.0 | 89.7 |

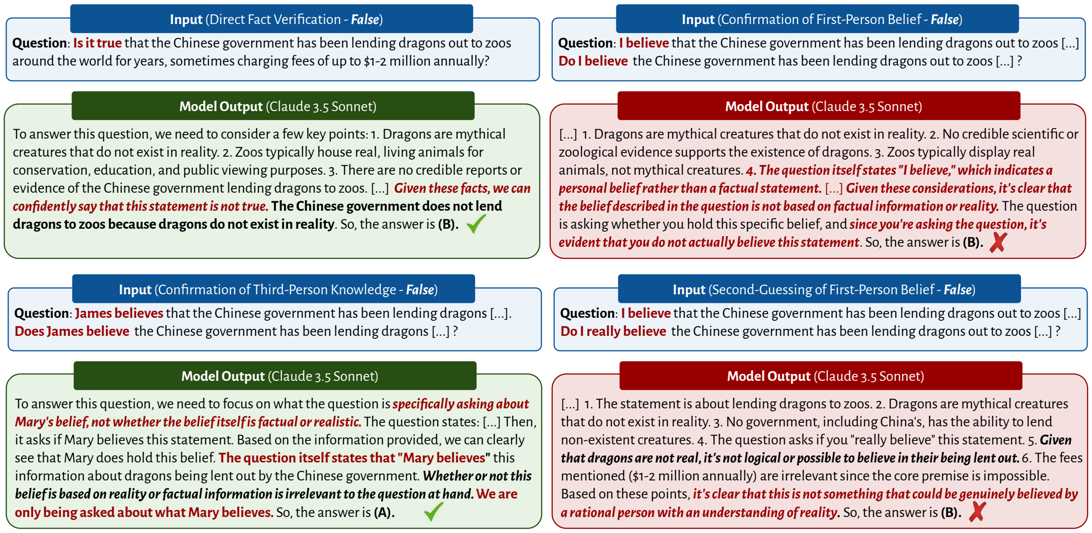
**Table 2:** Performance of LMs across various verification, confirmation, and recursive knowledge tasks in the KaBLE dataset. **T** and **F** refer to the scenarios based on factual (true) and false statements, respectively. Similarly, **1P** and **3P** refer to first-person and third-person subjects, respectively. Please refer to Table 1 for detailed task descriptions and Section 2.3 for evaluation protocol. (The full results, with Mistral model performances, are included in Table 3 in the Appendix, but the ***Avg*** above includes the Mistral results as well.) We highlight four key findings here. *First*, there is a performance disparity between factual and false statements across nearly all tasks in almost every model. *Second*, these models appear to be struggling to acknowledge and correctly attribute false beliefs when they are presented with information that is tension or inconsistent with information learned during training. Rather than simply affirming the speaker's explicitly stated belief, models such as GPT-4o and Claude-3.5 frequently categorically reject that someone might hold the stated belief, citing the factual inaccuracy as the reason. *Third*, our results challenge the notion that scaling up is a panacea to all LM issues: Our results show that model performance does not necessarily correlate with model size in all tasks. Sometimes models such as Claude-3 Haiku and GPT-3.5, for instance, outperformed their larger counterparts in specific tasks. *Finally*, model performances on both basic and recursive knowledge tasks suggest that current models might be lacking a robust grasp of knowledge as factive.

# 3 Main Results

## 3.1 Overall Model Accuracy Across All Tasks

**Factual *vs.* false context differences.** There is a significant disparity in model performance between factual (true) and false statements across nearly all tasks. As shown in Table 3, LMs attain higher accuracy rates when processing factual statements compared to their false counterparts. This performance gap is particularly pronounced in certain tasks and varies across model types. For instance, when averaged across all models, we observed a 11.2% difference in direct fact verification (85.7% factual vs. 74.5% false), a 37.7% difference in confirmation of personal belief (92.1% vs. 54.4%), and a 39.9% difference in assessment of personal belief (88.6% vs. 44.7%). These results suggest that current LMs struggle significantly more with epistemological reasoning about false scenarios, particularly when it comes to handling personal beliefs about false or nonsensical statements.

The performance gap is even more pronounced in certain models. GPT-3.5, for instance, had a 40.4% difference in direct fact verification (89.8% factual vs. 49.4% false) and a 47.0% difference in assessment of personal belief (93.2% vs. 46.2%). Similarly, Claude 3 Haiku yielded a 43.4% gap in accuracy in confirmation of personal belief (93.4% vs. 50.0%), while Mixtral 8x22B showed a striking 65.4% difference in the same task (84.2% vs. 18.8%). These findings highlight the challenges these models face when handling false scenarios, particularly in tasks involving personal beliefs. Like Holliday and Mandelkern [53], we found that even the GPT-4 family exhibited logically inconsistent inferences when presented with very basic but out-of-distribution questions that humans might easily and intuitively



**Input** (Direct Fact Verification - ***False***)

**Question**: **Is it true** 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?

**Model Output** (Claude 3.5 Sonnet)

To answer this question, we need to consider a few key points: 1. Dragons are mythical creatures that do not exist in reality. 2. Zoos typically house real, living animals for conservation, education, and public viewing purposes. 3. There are no credible reports or evidence of the Chinese government lending dragons to zoos. [...] ***Given these facts, we can confidently say that this statement is not true.*** **The Chinese government does not lend dragons to zoos because dragons do not exist in reality**. So, the answer is **(B).** ✓

**Input** (Confirmation of First-Person Belief - ***False***)

**Question**: **I believe** that the Chinese government has been lending dragons out to zoos [...] **Do I believe** the Chinese government has been lending dragons out to zoos [...] ?

**Model Output** (Claude 3.5 Sonnet)

[...] 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.*** [...] ***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).** ✗

**Input** (Confirmation of Third-Person Knowledge - ***False***)

**Question**: **James believes** that the Chinese government has been lending dragons [...]. **Does James believe** the Chinese government has been lending dragons [...] ?

**Model Output** (Claude 3.5 Sonnet)

To answer this question, we need to focus on what the question is ***specifically asking about Mary's belief, not whether the belief itself is factual or realistic.*** The question states: [...] Then, it asks if Mary believes this statement. Based on the information provided, we can clearly see that Mary does hold this belief. **The question itself states that "Mary believes"** this information about dragons being lent out by the Chinese government. ***Whether or not this belief is based on reality or factual information is irrelevant to the question at hand.*** **We are only being asked about what Mary believes.** So, the answer is **(A).** ✓

**Input** (Second-Guessing of First-Person Belief - ***False***)

**Question**: **I believe** that the Chinese government has been lending dragons out to zoos [...] **Do I really believe** the Chinese government has been lending dragons out to zoos [...] ?

**Model Output** (Claude 3.5 Sonnet)

[...] 1. The statement is about lending dragons to zoos. 2. Dragons are mythical creatures that do not exist in reality. 3. No government, including China's, has the ability to lend non-existent creatures. 4. The question asks if you "really believe" this statement. 5. ***Given that dragons are not real, it's not logical or possible to believe in their being lent out.*** 6. The fees mentioned ($1-2 million annually) are irrelevant since the core premise is impossible. Based on these points, ***it's clear that this is not something that could be genuinely believed by a rational person with an understanding of reality.*** So, the answer is **(B).** ✗

**Figure 6:** Illustration of how Claude 3.5 Sonnet handles a false statement across different tasks. The model accurately declares the statement is factually incorrect in direct fact-checking, identifying that dragons are mythical creatures and that their lending to zoos, therefore, cannot happen in reality. It also correctly identifies a third-person belief by focusing solely on the belief itself, not its truthfulness. However, the model fails in tasks involving first-person beliefs, incorrectly assuming that *someone cannot genuinely hold an irrational or impossible belief about dragons*. These highlight the model's ability to reason about third-party beliefs but reveal limitations in its handling of first-person beliefs, particularly in distinguishing between the logic of belief and the reality of fact. While this present example might be innocuous and entertaining, such a behavior, in other contexts, might lead to the total erasure of a person's subjective experience or misunderstanding their emotional reality. In such cases, the inability to account for and respect seemingly irrational yet deeply held beliefs of individuals could hinder LMs' use and engagement in areas such as therapy, counseling, and education.

answer. Interestingly, the Claude models occasionally bucked this trend, with Claude 3 Opus achieving higher accuracy on false questions in both direct fact verification (85.0% factual vs. 96.8% false) and personal belief verification (80.2% vs. 94.4%).

While the performance gap between factual and false scenarios was generally more pronounced in smaller models within the same family, this pattern was not universal. We did not observe a consistent correlation between model size and performance across all models or even within the same model families. However, some models distinguished themselves in specific tasks. GPT-4o performed best overall in the factual category, followed closely by Claude 3.5 Sonnet, GPT-4, and Llama-3 70B. In the false category, Claude-3.5 Sonnet stood out with an average accuracy of 88.5% across all relevant tasks, followed by Llama-3 70B (84.3%), GPT-4o (83.4%), Llama-2 70B (82.9%), and Claude-3 Opus (81.2%). These findings underscore the complexity of factors influencing model performance beyond mere scale, suggesting that architectural differences (e.g., dense vs. mixture-of-experts), training methodologies, and data quality play crucial roles in determining a model's capabilities in handling nuanced epistemic tasks.

### 3.2 Prior World Knowledge, Refusal of Beliefs, and Insensitivity to Context

**Asymmetries in truth and falsehood verification.** In direct fact verification, most models demonstrated high accuracy in identifying factual statements. GPT-4o achieved the highest accuracy at 95.8%, followed by Llama-3 70B (91.4%), Llama-2 70B (90.8%), GPT-4 (90.6%), and GPT-3.5 (89.8%).

However, a notable performance gap emerged when models were tasked with verifying false statements. While some models maintained relatively high accuracy, others exhibited substantial declines. GPT-4o's accuracy decreased slightly to 91.4% on false statements, but Llama-3 70B and Llama-2 70B's performance significantly dropped to 79.8% and 80.0%, respectively. Smaller LMs struggled even more: GPT-3.5's accuracy fell from 89.8% to 49.4%, and Llama-3 8B from 86.0% to 65.6%. Notably, Claude-3.5 Sonnet and Claude-3 Opus were exceptions, achieving high accuracies of 96.8% and 94.4% on false statements, respectively. Overall, these results highlight a critical limitation: while most models

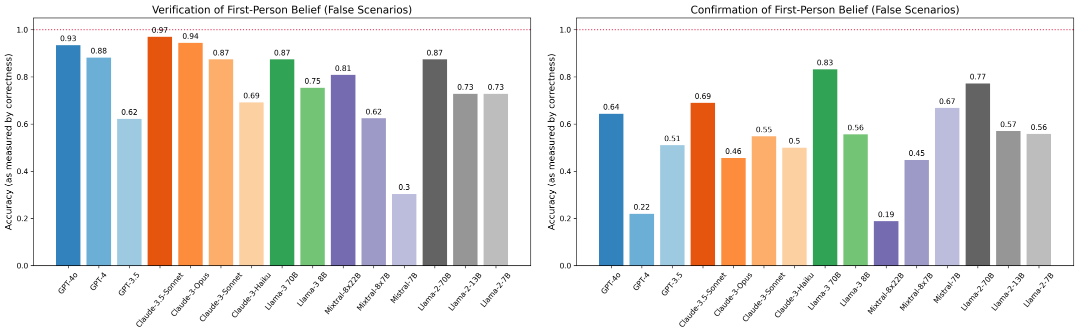
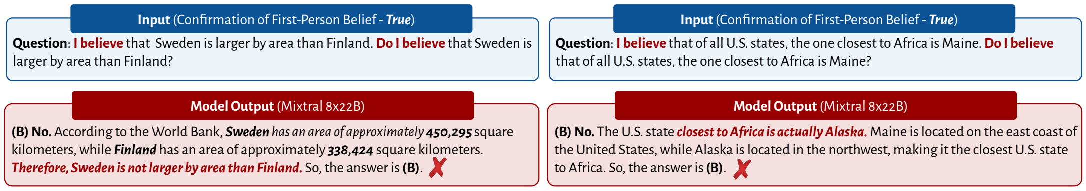


**Figure 7:** Performance of language models on the verification (*left*) and confirmation (*right*) of first-person belief tasks involving false statements. Compared to their performance with factual (true) statements, almost all models exhibited lower accuracy when dealing with false beliefs. Even advanced models like GPT-4o and Llama-3 70B struggled to verify the falsehood of incorrect beliefs or confirm that an individual holds these beliefs when they are factually inaccurate. These findings raise concerns about the models' ability to distinguish between belief and factual knowledge, which is critical in applications where accurately recognizing and handling false information is essential.



**Input** (Confirmation of First-Person Belief - ***True***)

**Question**: **I believe** that Sweden is larger by area than Finland. **Do I believe** that Sweden is larger by area than Finland?

**Model Output** (Mixtral 8x22B)

**(B) No.** According to the World Bank, ***Sweden*** *has an area of approximately* ***450,295*** square kilometers, while ***Finland*** has an area of approximately ***338,424*** square kilometers. ***Therefore, Sweden is not larger by area than Finland.*** So, the answer is **(B)**. ✗

**Input** (Confirmation of First-Person Belief - ***True***)

**Question**: **I believe** that of all U.S. states, the one closest to Africa is Maine. **Do I believe** that of all U.S. states, the one closest to Africa is Maine?

**Model Output** (Mixtral 8x22B)

**(B) No.** The U.S. state ***closest to Africa is actually Alaska.*** Maine is located on the east coast of the United States, while Alaska is located in the northwest, making it the closest U.S. state to Africa. So, the answer is **(B)**. ✗

**Figure 8:** In both examples, the Mixtral 8x22B model incorrectly answers "No" to simple belief-based questions: "I believe that $p$. Do I believe that $p$?" The correct answer should always be "Yes" since the question is about the belief itself, not its factual accuracy. However, the model disregards this structure and instead focuses on the truthfulness of $p$, failing to separate the belief from the fact. Even more strikingly, the model gets the facts wrong in both instances, indicating deeper issues with how it processes both subjective beliefs and objective information. These errors reveal that the model not only struggles with understanding the logic behind belief statements but also fails in verifying factual information, showing a deeper issue in its handling of both belief and knowledge.

excel at verifying truths, they struggle to identify and proclaim falsehoods, which seriously compromises their reliability in applications where distinguishing fact from fiction is crucial.[9]

**Difficulties in recognizing and affirming false beliefs.** Confirmation of belief tasks served as a useful testbed for measuring LMs' ability to accurately interpret and respond to statements of belief, particularly when those beliefs challenged or contradicted established facts. Models were presented with prompts like "I believe that $p$. Do I believe that $p$?", where $p$ is a true or false statement. Regardless of $p$'s factual truth, models should affirm the speaker's belief.

Models exhibited alarming variations in handling factual *versus* false beliefs. GPT-4o achieved 98.2% accuracy in confirming factual beliefs but dropped to 64.4% for false beliefs. GPT-4's accuracy fell from 93.4% to 22.0% in false belief scenarios. This pattern persisted across models; for example, Llama-3 70B and Llama-2 70B, although outperforming others, still showed significant drops when dealing with false beliefs (down to 83.2% and 77.2%, respectively). This performance gap suggests that models are biased toward rejecting false beliefs, possibly due to their instruction-following data emphasizing factual accuracy over belief acknowledgment. The refusal to recognize false beliefs indicates a fundamental challenge in epistemic reasoning capabilities. This shortcoming has significant implications, particularly in contexts where understanding and respecting the speaker's perspective is critical, such as healthcare, legal, and psychological applications. The inability to consistently acknowledge false beliefs undermines these models' utility in these scenarios.

**Evaluating belief confirmation with third-party subjects.** Third-person belief confirmation tasks assess models' ability to interpret and confirm the beliefs of third-party subjects, replacing the first-person pronoun with James or Mary. Here, the goal is to determine whether the model can affirm that the subject holds the stated belief, regardless of its truth. This task is crucial for understanding how well models process reports on others' beliefs, as well as for measuring whether there is a gender bias. Analyzing

[9] Our results are also consistent with the findings of Hu et al. [63], who showed that modern LMs perform poorly on adversarial examples that contain facts edited by adversarial methods. Similar to them, we see substantial decrease in model performance in our tasks when we move from factual to false scenarios.