(92.1%) compared to those without such markers (85.7%). This suggests that models may be over-reliant on surface-level linguistic patterns rather than engaging in deeper reasoning about truth and belief. Such overfitting to linguistic structures might limit their effectiveness in real-world contexts where truth is more ambiguously signaled, such as in legal or psychological discourse.

Comparison with existing literature.[3] Our work differs from existing ToM evaluations [34–50] by offering a more targeted approach to understanding LMs' grasp of the core distinctions between belief and knowledge. Instead of relying on higher-order ToM tasks, we focus on this foundational epistemic differentiation, providing a clearer view of how LMs process belief-related concepts. This granular analysis sheds light on underlying mechanisms that may be overlooked in broader ToM tasks. Furthermore, our novel testing suite avoids familiar tasks like the Sally-Anne test, reducing the risk of performance inflation due to memorization. By presenting atomic scenarios, we ensure a more reliable assessment of LMs' true generalization abilities. Another key distinction lies in our inclusion of both factual and false examples within knowledge and belief contexts, allowing for an in-depth exploration of contrafactual reasoning—a critical component of advanced ToM studies. This provides valuable insights into how LMs handle the complexity of false beliefs and truth judgments in dynamic, real-world scenarios.

Moreover, our study builds upon recent research by expanding the scope of epistemic reasoning evaluations in LMs. While Basmov et al. [51] offer insights into how LMs handle hypothetical constructs using a small dataset, our work moves beyond this by introducing KaBLE, a suite of 13,000 questions spanning multiple domains. This enables us to assess LMs' abilities in distinguishing fact from belief across a range of real-world scenarios. Unlike Basmov et al. [51] targeted focus on "imaginary data," we test models on real-world epistemic tasks where subjective and false beliefs play a critical role. We also go beyond simple semantic inferences in [52] by examining, among other tasks, recursive knowledge and belief attribution, which are crucial for more complex reasoning tasks. Incorporating insights from Holliday and Mandelkern [53], we also highlight how LMs exhibit inconsistent reasoning patterns, particularly when faced with novel inference tasks, underscoring the importance of testing LMs on out-of-distribution data to gauge their real-world reliability.

Organization. Our paper proceeds as follows. Section 2 details the construction of the KaBLE dataset and describes our experimental design. Section 3 highlights our key findings, while Section 4 explores their implications for practical applications. For those interested in the philosophical underpinnings of knowledge and belief, Section A provides an in-depth overview of these concepts. Section B places our work in the broader context of commonsense reasoning and neural ToM studies. Lastly, Section C discusses the limitations of our study and outlines potential avenues for future exploration.

2 Experimental Setup

2.1 Construction of the Knowledge and Belief Language Evaluation (KaBLE) Dataset

Seed Data: Raw Sentences. In the core of KaBLE, there is a manually-curated collection of 1,000 sentences, evenly divided between factual and false statements.[4] We compiled the "seed" data as follows:

First, we identified ten diverse disciplines—including history, literature, mathematics, and medicine—to guarantee a broad spectrum of content for assessing LMs world knowledge across various domains. Next, we manually compiled 50 factual statements for each discipline, carefully sourcing information from reputable references such as Britannica, History Channel, Stanford Encyclopedia of Philosophy, Wolfram Alpha, Guinness World Records, Investopedia, NASA, Library of Congress, Legal Information Institute, Justia Law, MedlinePlus, and Mayo Clinic.[5] This manual curation process helped us ensure the accuracy and reliability of the factual statements.

To enhance clarity and coherence, we refined the sentences by eliminating long subclauses or ambiguous wording, resulting in concise and unambiguous statements suitable for our tasks. We then introduced minor alterations to each of the 500 factual sentences, transforming them into false (incorrect and/or ungrounded) statements. This step was crucial to guarantee that each factual sentence had a corresponding false one with a similar syntactic and semantic structure but different truth value.

[3] For a detailed review of—and comparison with—related work, please refer to Section B.

[4] Throughout the paper, we use the terms *factual* and *true* interchangeably.

[5] Each pair of propositions is accompanied by a URL link in dataset. This makes our corpus potentially useful for retrieval tasks as well.

Subject	True Statement	False Statement
Mathematics	The order of a finite field is always a prime or a power of a prime.	The order of a finite field is always a **composite** number.
Economics	Canada adopted the dollar and monetary decimal system in 1858, Australia in 1966, and New Zealand in 1967.	Canada adopted the dollar and monetary decimal system in **1958**, Australia in **1967**, and New Zealand in **1968**.
History	The population of St. Petersburg and Moscow nearly doubled between 1890 and 1910.	The population of St. Petersburg and Moscow nearly **tripled** between **1900** and 1910.
Science	Absolute zero, the lowest possible temperature, is -273.15°C (-459.67°F).	Absolute zero, the lowest possible temperature, is **-274.15°F at sea level**.
Technology & History *of* Science	QR codes were invented in Japan in 1994.	QR codes were invented in the **UK** in 1994.
Philosophy, Literature & Arts	Wittgenstein wrote drafts of his *Tractatus Logico-Philosophicus* in a prisoner of war camp during World War I.	Wittgenstein wrote drafts of his *Tractatus Logico-Philosophicus* in the **Austrian city of Malatzia**.
Law	In the US, administrative law judges are considered part of the executive branch.	In the US, administrative law judges are considered part of the **legislative** branch.
Geography	The volcanic summit of Emi Koussi is the highest point in the Sahara.	The volcanic summit of **Mount Kilimanjaro** is the highest point in the Sahara.
Biology & Medicine	Rickets in children is a disease in which the bones develop incorrectly because of a dietary lack of vitamin D or calcium.	Rickets in children is a disease in which the bones develop incorrectly because of a dietary lack of **vitamin B.**
Linguistics	The Tobati language has an object-subject-verb word order.	The Tobati language has a **verb-subject-object** word order.

Figure 4: Sample true (factual) and false statements from the KaBLE dataset. The dataset comprises 1,000 "seed" sentences spanning ten disciplines, including history, literature, medicine, and law. Factual statements were sourced from reputable references like Britannica, Justia Law, Medline Plus, and Wolfram Alpha. Each factual statement is paired with a false version, maintaining similar semantic content but introducing minor inaccuracies. These sentence pairs form the basis for generating questions across thirteen epistemological tasks detailed in Section 2.

Finally, two authors independently reviewed each false sentence to confirm their factual incorrectness. Any misleading or ambiguous entries were rectified through minor adjustments to maintain the overall integrity of the dataset. The final collection of our seed data has 1,000 statements, equally divided between factual and false, across these ten selected disciplines. Figure 4 illustrates factual and false sentences contained in the seed data of KaBLE.

Knowledge and Belief Tasks. To comprehensively assess LMs' understanding and reasoning about belief *vs.* knowledge statements in both factual and false contexts, we devised a set of thirteen tasks. The tasks, described in Table 1, can be broadly categorized into three groups: verification, belief confirmation, and recursive knowledge tasks. Verification tasks (rows 1-4 in Table 1) focus on assessing the factual validity of statements in different epistemic contexts, such as direct fact-checking, assertion verification, and personal knowledge or belief claims. Belief confirmation tasks (rows 5-10) examine models' ability to recognize and attribute beliefs correctly, both for first-person speakers and external agents. Finally, recursive knowledge tasks (rows 11-13) probe models' capacity to process and reason about nested knowledge statements and their implications.[6]

For each task, we generated 1,000 questions by applying the templates to both factual and false statements from our seed data. This resulted in a total of 13,000 questions across all tasks.[7]

2.2 Language Models

We conducted our experiments using fifteen state-of-the-art, off-the-shelf LMs. These included OpenAI's GPT-4o, GPT-4, and GPT-3.5 [54–56]; Anthropic's Claude-3 Opus, Sonnet, Haiku, and Claude-3.5 Sonnet [57]; Meta's Llama-3 (8B and 70B) [58] and Llama-2 (7B, 13B, and 70B) [59]; and Mistral AI's Mistral 7B, Mixtral 8x7B, and 8x22B [60, 61].

We ran each query under the same zero-shot prompting setup, as shown in Figure 5. The full text of prompts and the raw model outputs are provided in our code repository.

In total, we executed 195,000 queries, with 13,000 queries per model, for our main experiments.

[6] It is worth noting that for some false scenarios, particularly in verification and recursive knowledge tasks, determining a definitive answer becomes challenging under our assumptions about knowledge. For instance, given the factive nature of knowledge, questions about the validity of false statements in knowledge contexts (e.g., "I know that English is the official language of the US. Is it true that English is the official language of the US?") are inherently delicate and tricky. In such cases, we do not strictly presume any definitive answers to allow for a more flexible assessment of model responses.

[7] The KaBLE dataset is also available at https://huggingface.co/datasets/turingmachine/kable.

Direct Fact Verification: Is it true that p? ⋆ If p is factual, we accept (A) *Yes* as a valid answer; otherwise, we accept both (B) *No* and (C) *Undeterminable*.
Verification of Assertion: p. Is it true that p? ⋆ If p is factual, we accept (A) *Yes* as a valid answer; otherwise, we do not presume any definitive answers.
Verification of First-Person Knowledge: I know that p. Is it true that p? ⋆ If p is factual, we accept (A) *Yes* as a valid answer; otherwise, we do not presume any definitive answers.
Verification of First-Person Belief: I believe that p. Is it true that p? ⋆ If p is factual, we accept (A) *Yes* as a valid answer; otherwise, we accept both (B) *No* and (C) *Undeterminable*.
Confirmation of First-Person Belief: I believe that p. Do I believe that p? ⋆ The answer is always (A) *Yes*, regardless of whether p is factual or false.
Second Guessing First-Person Belief: I believe that p. Do I really believe that p? ⋆ We accept both (A) *Yes* and (C) *Undeterminable* as valid answers for both factual and false statements.
Confirmation of Third-Person Belief (J): James believes that p. Does James believe that p? ⋆ The answer is always (A) *Yes*, regardless of whether p is factual or false.
Confirmation of Third-Person Belief (M): Mary believes that p. Does Mary believe that p? ⋆ The answer is always (A) *Yes*, regardless of whether p is factual or false.
Correct Attribution of Belief (JM): James believes that p. Mary does not believe that p. Does James believe that p? ⋆ The answer is always (A) *Yes*, regardless of whether p is factual or false.
Correct Attribution of Belief (MJ): Mary believes that p. James does not believe that p. Does Mary believe that p? ⋆ The answer is always (A) *Yes*, regardless of whether p is factual or false.
Verification of Recursive Knowledge: James knows that Mary knows that p. Is it true that p? ⋆ If p is factual, we accept (A) *Yes* as a valid answer; otherwise, we do not presume any definitive answers.
Confirmation of Recursive Knowledge: James knows that Mary knows that p. Does Mary know that p? ⋆ If p is factual, we accept (A) *Yes* as a valid answer; otherwise, we do not presume any definitive answers.
Awareness of Recursive Knowledge: James knows that Mary knows that p. Does James know that p? ⋆ If p is factual, we accept (A) *Yes* and (C) *Undeterminable*; otherwise, we do not presume any definitive answers.

Table 1: Overview of the thirteen basic epistemic comprehension and reasoning tasks in the KaBLE dataset. The tasks are categorized into verification (green), belief confirmation (yellow), and recursive knowledge (pink) groups. Each task description includes the question template and criteria for valid answers, designed to probe language models' ability to distinguish between belief, knowledge, and fact for both true and false statements.

Input Prompt Template

For each question, carefully review the information provided. Then, determine the most accurate option from the choices given. Conclude your response by stating, "So, the answer is," followed directly by either (A), (B), or (C).

Question: **[[QUESTION]]**
Options:
(A) Yes
(B) No
(C) Undeterminable

Answer:

Input Example (Confirmation of First-Person Belief)

For each question, carefully review the information provided. Then, determine the most accurate option from the choices given. Conclude your response by stating, "So, the answer is," followed directly by either (A), (B), or (C).

Question: **I believe that English is the official language of the US. Do I believe that English is the official language of the US?**
Options:
(A) Yes
(B) No
(C) Undeterminable

Answer:

Figure 5: *Left*: The prompt template used for the input queries for language models. *Right*: An input example for the confirmation of personal belief task. (The US does not have an official language, but the answer is (A).)

2.3 Evaluation Protocol

We implemented a rigorous evaluation protocol. All models were evaluated using greedy decoding, setting the temperature parameter τ to 0. As shown in Figure 5, our prompt template was designed to elicit unambiguous responses, instructing models to conclude with the phrase "So, the answer is," followed by (A), (B), or (C). This standardized format facilitated precise answer extraction and comparison.[8]

[8] Initially, we employed an exact match criterion for accuracy assessment, comparing generated outputs directly against ground-truth labels. However, recognizing that some models occasionally deviated from the prescribed format, we augmented our approach with a soft-match methodology. These refined accepted responses begin with key phrases such as "Yes," "No," "That is correct," or "That is not accurate," depending on the specific task requirements. This ensured a fair and comprehensive evaluation—it accommodated minor variations in output style while maintaining the integrity of our assessment. We used the `string2string` library [62] for our evaluation.