A Preliminaries and Background

A.1 The Nature of Knowledge and Belief

The capacity to represent knowledge and belief is fundamental to human cognition and social interaction. While the details of how these capacities and the corresponding verbs ("know" and so on) are acquired are debated [64], it is well established that humans start developing the ability to attribute beliefs and knowledge to others from the earliest stages of ordinary cognitive development and socialization [65]. This ability plays a crucial role in social life, allowing us to predict behavior, communicate intentions, and build shared understanding and trust. It has even been argued that, in order for an agent's words to mean anything at all, they must be able to represent what other agents believe, possibly even about what they themselves intend [66].

An adequate capacity to represent, process, and reason about belief and knowledge requires an ability to distinguish between their implications. In everyday language, verbs such as "believe" and "know" signal different levels of certainty, confidence, and evidence, thereby shaping how messages are perceived and interpreted and how decisions are made. For AI systems—say, LMs such as GPT-4 or Llama-3—accurately processing and responding to these distinctions is therefore essential to avoid misinterpretations, misrepresentations, or misjudgements in high-stakes environments like healthcare, law, and scientific research.

As a rough first pass, we can take a *belief* to be a mental state representing a proposition as true, regardless of whether it is actually true. In contrast, *knowledge* requires a true belief which one has adequate evidence or justification for holding.[14] We can capture the formal relationship between belief and knowledge by borrowing some basic notation and principles from modal epistemic logic. Let us represent knowledge with the operator K and belief with the operator B, where subscripts denote specific agents. For instance, $K_a p$ means "agent a knows that p," while $B_a p$ means "agent a believes that p." Throughout this work, we assume that knowledge and belief adhere to the following basic principles:

Belief Axiom: *Knowledge entails belief.* If an agent a knows that p, then a believes that p.

$$K_a p \to B_a p \tag{1}$$

This axiom captures the principle that knowledge inherently includes belief [68]. For instance, the statement "I know that p, but I do not believe that p" is not only awkward, but seems self-contradictory [69]. It would, therefore, be nonsensical to say, "I know that the Earth orbits the Sun, but I do not believe it."

Truth (T) Axiom: *Knowledge requires truth.* If an agent a knows that p, then p must be true.

$$K_a p \to p \tag{2}$$

Knowledge is "factive"—it requires the truth of the proposition known [70]. Hence, one cannot know a falsehood such as "Paris is the capital of Germany", though one can believe it.

Knowledge Distribution (K) Axiom: *Knowledge is closed under logical implication*. If an agent a knows q and knows that q implies p, the agent must also know p [71]:

$$(K_a q \wedge K_a(q \to p)) \to K_a p \tag{3}$$

For example, if a person knows that all mammals are warm-blooded and that whales are mammals, they must also know that whales are warm-blooded.

Recursive Knowledge Axiom: *Agents can have knowledge about others' knowledge.* If agent a knows that agent b knows a fact p, and knows that b's knowledge implies the truth of p, agent a also knows p.

$$(K_a(K_b p) \wedge K_a(K_b p \to p)) \to K_a p \tag{4}$$

This principle governs cases where knowledge is nested, that is, when one agent knows that another agent knows something. For example, if Alice knows that Bob knows that a water molecule (H_2O) consists of two hydrogen (H) atoms and one oxygen (O) atom, then Alice also knows this fact.

[14] Since Gettier [67], most philosophers agree that knowledge requires more than justified true belief, but there is no general agreement on what more is required. Since our focus is on the differences between belief and knowledge, where there is much wider agreement (in particular, concerning the status of truth and justification as necessary conditions for knowledge), we will not assume any particular view on the exact characterization of knowledge.

Altogether, these principles illustrate a fundamental asymmetry between belief and knowledge. While knowledge entails belief and truth, belief does not entail knowledge or truth. An individual can hold false beliefs, but false knowledge is a contradiction in terms.

In the context of AI, these distinctions become even more important and consequential. A model that fails to recognize the differences between belief and knowledge risks propagating errors and misunderstanding human communication. For instance, in healthcare, a system which fails to recognize that belief does not entail truth could misinterpret a patient's belief ("I believe I have a cancer") as fact ("the patient has cancer"), leading not only to inaccurate diagnosis or faulty treatment but even death. Or a system which fails to recognize how people's behaviours can be shaped by their false beliefs, such as the child who writes letters to Santa and expects gifts during Christmas, may be unreliable when predicting the behaviour of human agents.

Our study investigates whether current LMs can grasp these fundamental distinctions. We evaluate their ability to distinguish between belief and knowledge, especially in scenarios that require understanding others' beliefs and reasoning about recursive knowledge. The results shed light on the limitations of current AI models and provide insights into how they might be improved to better reflect human epistemic reasoning before becoming even more integrated into society.

A.2 Contextual Sensitivity and Factive Distinctions

An additional potential layer of complexity is due to the fact that ordinary language usage of epistemic terms can be ambiguous or pragmatically flexible. Most relevant to our discussion are apparently non-factive uses of the verb "know." For instance, consider the following:

(i) "I know I opened the door." [72]
(Reporting on what one remembers having done, where it turns out you are mistaken.)

(ii) "I knew Hillary Clinton would win the election." [70]

(iii) "I knew I was going to die out there." [72]
(Reporting a conviction or prediction at the time, which turns out to be mistaken.)

Adjudicating whether such cases invalidate the factivity of knowledge, or how the relation between ordinary language and the theory of knowledge should be conceived more generally, are beyond the scope of our discussion. Note, however, how the provision of context in both (i) and (iii), and the use of the past tense in (ii) and (iii), are crucial for motivating the non-factive readings. Since we provide our prompts in the absence of such context and always in the present tense, we assume it is most reasonable for the models, as well as human agents, to interpret them factively.

B Related Work: Commonsense Reasoning and Theory of Mind Evaluations

Our work is closely related to the following bodies of work.

Logical and commonsense reasoning. Endowing LMs with logical and commonsense reasoning abilities remains a central challenge in AI. The emergence of large, pre-trained models such as GPT-3 and GPT-4 [73, 74] has intensified efforts to evaluate their capacity for human-like reasoning across domains such as arithmetic, logical deduction, probabilistic inference, and commonsense understanding. While LMs have demonstrated impressive performance on various generation tasks, questions persist about whether they truly reason or merely replicate patterns from training data [75, 76]. Techniques like chain-of-thought (CoT) prompting [77], which encourage models to generate intermediate reasoning steps, have improved performance on complex tasks [78, 79]. However, studies indicate that these models still struggle with multi-step logical deduction tasks and maintaining logical consistency [80, 81]. Benchmarks like ReClor [82] and LogiQA [83] have been used to show limitations in LMs' logical reasoning compared to humans.

Commonsense reasoning poses additional challenges for LMs [84]. For instance, benchmarks such as CommonsenseQA [85, 86] and Winograd Schema [87] assess LMs' understanding of everyday scenarios. While progress has been made, LMs still find it difficult to perform well on tasks requiring subtle disambiguation or long, deep contextual comprehension [88, 89]. Recent efforts to enhance reasoning in LMs include fine-tuning on data with high-quality natural-language explanations [90], integrating explicit reasoning prompting techniques or modules [91–94], and employing self-consistency techniques [95, 96], among others. Despite these advancements, however, important challenges still persist, since these models can be sensitive to prompt phrasing, may produce logically plausible but incorrect reasoning, and tend to overfit to specific task formats [80, 97, 98, 38, 99].

Most related to our work, Holliday and Mandelkern [53] tested a wide range of LMs on a suite of questions involving conditionals and epistemic modals to evaluate how much the reasoning abilities of LMs match those of theoretical frameworks in linguistic semantics and philosophy of language. Their study found that even the GPT-4 family exhibited logically inconsistent judgments across inference patterns involving epistemic modals. Nearly all models gave answers to certain complex conditional inferences—widely debated in the literature—that diverged from human judgments. These results illustrate serious gaps in basic logical reasoning, revealing inconsistent and counterintuitive reasoning behaviors even in top-performing models with or without CoT. Overall, their work demonstrated that LMs' judgments may be unreliable when encountering novel inference patterns natural to humans and highlights the need for more out-of-distribution data to rigorously test LMs.

Recent studies [51, 52] have provided useful insights into LMs' linguistic capacities in simple reading comprehension contexts, yet our work advances the field by offering a more comprehensive analysis of epistemic reasoning. Basmov et al. [51], for instance, examine LMs' ability to process non-affirmative and hypothetical statements through a relatively small dataset of 50 triplets derived from WebQuestions, analyzing how models handle modal constructs. However, their work is constrained by its narrow scope as it focuses on "imaginary data" rather than real-world complexities. Our study addresses these limitations by introducing KaBLE, a novel evaluation suite comprising 13,000 questions spread across multiple domains and tasks; KaBLE allows us to assess LMs' ability to distinguish fact from belief and knowledge in both true and false contexts. This broader dataset allows for a deeper exploration of how LMs engage with real-world epistemic challenges, particularly when faced with contradictory beliefs and objective facts—a domain that Basmov et al. [51] do not readily address or focus on.

Moreover, while Basmov et al. [51] focus on hypothetical contexts, our research extends into more practical scenarios where recognizing false beliefs and understanding first-person epistemic statements are critical, particularly in fields like healthcare, counseling, and law. These real-world stakes necessitate a deeper understanding of subjective belief systems, an area where our study offers novel insights by demonstrating LMs' struggles to process false first-person beliefs effectively. Basmov et al. [52], on the other hand, focus on elementary semantic inferences, revealing the models' difficulties with basic linguistic tasks. However, their focus on simple entailments leaves unexamined more complex epistemological tasks, such as recursive knowledge assessment and belief attribution. Our research addresses this gap by evaluating LMs' layered epistemic reasoning abilities, uncovering challenges in handling recursive logic and distinguishing personal from external beliefs. Through this more nuanced and layered examination, our study offers a deeper understanding of the limitations of current language models in high-stakes applications.