Theory of mind evaluations on modern LMs. Theory of mind (ToM) refers to the capacity to process, explain, and predict people's observable actions in terms of their mental states, such as beliefs, desires, and intentions [100]. ToM is considered to be a foundational aspect of human cognitive development and social interaction [101–103]. Recently, as modern LMs have demonstrated increasingly sophisticated language understanding and generation abilities, researchers have begun to investigate whether these models can understand and reason about mental states similar to how humans do. Several recent studies [34, 35] found that models such as GPT-3 performed poorly on simple ToM tasks in zero-shot settings. Subsequent studies [39, 36], however, have claimed that more recent and larger models such as GPT-4 display emergent ToM capabilities. Bubeck et al. have even gone as far as to claim that GPT-4 has shown "sparks of artificial general intelligence" and that their "findings suggest that GPT-4 has a very advanced level of theory of mind."

These disparate findings and bold claims sparked a lively debate in the AI community and were quickly met with skepticism and critical examinations [40–42, *inter alia*]. Marcus and Davis [104], for instance, outlined many methodological issues with the experiments conducted by Kosinski [36], while Ullman [37] and Shapira et al. [38] demonstrated that minor alterations to test samples could dramatically reduce model performance on elementary ToM tasks. Shapira et al. [38] further argued that these large-scale LMs might be relying on shallow heuristics and superficial patterns rather than displaying genuine understanding of these ToM tasks.

ToM and social reasoning benchmarks. The recent challenges in accurately assessing ToM capabilities of LMs have spurred the development of more comprehensive benchmarks and evaluation methods. Evaluation suites such as ToMi [43], ToMChallenges [44], BigToM [45], HiToM [46], EPITOME [47], Thinking for Doing [T4D; 48], OpenToM [49], *inter alia*, have provided valuable tools for probing different aspects of ToM reasoning. However, as highlighted by Ma et al. [105] and others [43, 41, 38], there is growing recognition in the field that existing benchmarks may still not be fully capturing the nuanced cognitive processes underlying ToM and that they are also prone to data contamination, spurious correlations, and implicit linguistic biases. This realization has led to calls for more holistic approaches to evaluation, including methods and benchmarks that can assess precursor abilities such as perception inference [106] and perception-to-belief inference [107, 108]—as exemplified in [109]—or provide a more comprehensive and conclusive picture of LM's capabilities and limitations in ToM and social reasoning tasks. Overall, as noted by Mahowald et al. [42], "[t]o assess progress on the road toward building models that use language in human-like ways, it is important to develop benchmarks that evaluate both formal and functional linguistic competence."

Comparison with existing ToM evaluations. Our work distinguishes itself from existing ToM evaluations in several critical ways. First, it provides a more direct and principled approach to assessing LMs' comprehension of belief *vs.* knowledge statements. By focusing on this fundamental epistemic distinction, we probe a crucial precursor to more complex ToM reasoning. This allows for a more granular analysis of LMs' capabilities, potentially revealing insights into their underlying mechanisms for processing epistemic concepts. Second, our work introduces a novel testing suite with an extensive quantity of original content, deliberately avoiding well-known and well-documented ToM tasks like the Sally-Anne test [110], which may have been encountered during training. This design choice mitigates the risk of performance inflation due to memorization or overfitting to specific task structures. Instead, it tests models under fresh and atomic scenarios, providing a more robust and simple assessment of their generalization capabilities in epistemic and social reasoning. Third, our work's inclusion of both factual and false examples in the context of knowledge and belief adds a layer of complexity absent in many existing evaluations. This feature allows for a nuanced exploration of how LMs handle truth values in relation to epistemic states, shedding light on their capacity for reasoning in "contrafactual" scenarios, which is crucial in studying advanced ToM abilities [see also: 50]. By addressing these aspects, our work thus seeks to provide a more comprehensive and illuminating picture of modern LMs' epistemic processing and social reasoning capabilities.

C Limitations and Future Directions

Despite the valuable insights offered by this study, several limitations must be acknowledged. These pertain primarily to the linguistic complexity of epistemic reasoning, the design of our experimental tasks, and the inherent challenges in evaluating LMs on epistemic distinctions. By recognizing these constraints, we can better understand the boundaries of our findings and identify areas for future research.

Linguistic complexity and pragmatic considerations. One of the central limitations of our study is the oversimplification of the linguistic nuances surrounding epistemic phrases such as "know" and "believe." While our experiments focused on a core set of epistemic terms, we did not sufficiently account for the flexibility and contextual fluidity with which these phrases are used in everyday language. In many colloquial contexts, the word "know" is used interchangeably with "believe" or "think." People often say, "I know" when they mean "I believe" or "I think," introducing ambiguity in the relationship between the speaker's epistemic stance and the literal truth of the statement.

Furthermore, epistemic utterances are not always straightforward in their intent. They can be performative, rhetorical, sarcastic, or satirical. For example, when someone says, "I know the sky is green," the statement may not reflect their genuine belief but could instead be a satirical comment meant to challenge an interlocutor's assumption. LMs, however, are often not equipped to detect these pragmatic subtleties, leading to erroneous conclusions about what constitutes belief, knowledge, or fact. This pragmatic gap limits the models' capacity to engage with language as it is used in real-world human interactions, and future studies should address this issue by incorporating more diverse linguistic contexts.

Specific focus on select epistemic phrases. Our study primarily focused on the epistemic phrases "know," "believe," and "think," overlooking other common epistemic expressions such as "I know," "I feel," "I suppose," "I gather," "I imagine," and "I presume." Each of these phrases conveys varying degrees of epistemic commitment and subjective belief, and their inclusion could provide a more comprehensive view of how LMs handle epistemic reasoning. By limiting our analysis to a small subset of epistemic phrases, we may have restricted our understanding of the models' broader capabilities in processing and distinguishing between different levels of certainty and belief. A more expansive exploration of the full range of epistemic expressions could offer deeper insights into how LMs manage different shades of epistemological reasoning, particularly in more complex, context-sensitive scenarios.

Influence of complementizer usage. Another linguistic issue relates to the presence or absence of the English complementizer "that" in the statements we presented to the models. Research has shown that the inclusion or omission of "that" can subtly influence the interpretation of epistemic authority [111]. For instance, the phrase "I know that the Earth orbits the Sun" may be interpreted differently from "I know the Earth orbits the Sun," with the latter potentially perceived as more assertive or informal. Although we attempted to standardize the structure of our statements, the decision to include the complementizer may have had some unintended effects on model performance. This variability in sentence structure could introduce noise into the results and confound the models' ability to accurately process epistemic claims. Future studies should more rigorously examine how different syntactic structures influence model reasoning in epistemic contexts.

Limited exploration of contextualism. Context plays a critical role in how epistemic statements are interpreted. In everyday conversations, the meaning of phrases like "I know" or "I believe" can change depending on the social, cultural, or conversational context in which they are uttered. However, our study did not fully account for the contextualism inherent in epistemic reasoning. LMs often rely heavily on the immediate textual input without fully integrating broader contextual cues that might inform the speaker's intent or epistemic state. For example, in legal or medical settings, "knowing" carries a far more stringent epistemic burden than in casual conversation. This lack of contextual sensitivity in our evaluation may have led to an oversimplified view of how LMs handle complex epistemic tasks. Addressing this issue would require more sophisticated, multi-modal analyses that incorporate not only linguistic inputs but also situational, social, and cultural cues that shape how epistemic phrases are understood.

Task design and model generalization. Our task design, while comprehensive, may have limited the generalizability of our findings. The tasks we employed, although carefully crafted to evaluate models' epistemic reasoning, were designed as isolated test cases, often lacking the rich, dynamic conversational flow that would be present in real-world scenarios. Consequently, the performance of LMs in these controlled settings may not directly translate to more complex, open-ended dialogues where epistemic distinctions are embedded in multi-turn conversations. This gap between task-based evaluation and real-world application remains a challenge for all current AI research and should be addressed in future studies by incorporating more interactive, dialog-based evaluations that better reflect the nuanced nature of human epistemic reasoning.

Impact of zero-shot CoT on performance. To explore ways to enhance models' epistemic reasoning, we applied the zero-shot CoT prompting method [91] to GPT-4o. This technique, which involves prefacing responses with "Let's think step by step," was designed to encourage the model to engage in more deliberate and structured reasoning. The outcomes of this intervention were mixed and somewhat inconclusive. In tasks focused on direct fact verification and first-person belief verification, the model's performance dropped by 6% and 4%, respectively, indicating that CoT prompting did not enhance reasoning in these areas. In contrast, this approach showed more promise in tasks involving complex reasoning. In the confirmation of first-person beliefs, especially in false scenarios, performance increased by 11%, leading to an overall 6% improvement in first-person belief reasoning. Additionally, a 10% performance boost was observed in factual scenarios related to recursive knowledge awareness, although recursive knowledge confirmation only saw a marginal 0.4% increase.

Despite these gains, the model continued to display some of the earlier weaknesses, such as its difficulty recognizing that individuals can hold false beliefs. These findings suggest that while zero-shot CoT prompting can enhance performance in certain complex, non-verification tasks, it is still not a panacea for the deeper epistemic limitations observed in LMs. More advanced strategies may be needed to fully address these challenges and improve the model's overall reasoning capabilities.