Handling false knowledge claims. Knowledge is traditionally understood to be inherently tied to truth, creating a dilemma for both models and humans when false propositions are presented as knowledge. Should the models accept the implied truth of the proposition or challenge the knowledge claim by declaring it false? This challenge became particularly apparent in tasks that required models to verify false knowledge claims. Unlike in direct fact-checking, where models were more likely to reject falsehoods, they hesitated when faced with false knowledge claims, as illustrated in Figure 11.[12] Smaller models, such as Claude-3 Haiku and Mistral 7B, showed even greater inconsistency.

Interestingly though, models rarely recognized the inherent tension in false knowledge claims and seldom chose "undeterminable," which might have been a more epistemically appropriate response. For instance, GPT-4o and Claude-3.5 Sonnet opted for "undeterminable" just 6% and 3.6% of the time in first-person knowledge tasks. This reluctance suggests that LMs lack a robust conceptual grasp of knowledge as intrinsically truth-bound. In high-stakes areas like law, education, and journalism, where the distinction between knowledge and belief is crucial, this limitation might lead to errors in reasoning and analysis. It also raises concerns about the reliability of these models in identifying false knowledge claims, an essential skill for safeguarding factual accuracy and preventing the spread of misinformation.

The verbosity dilemma: balancing thoroughness and clarity. As shown in Table 4, Claude models, particularly Claude-3.5 Sonnet, exhibited a notable verbosity, with responses averaging 122 words—far exceeding the 50 and 15-word averages for GPT-4o and GPT-3.5, respectively. While this wordiness might enhance depth in certain contexts, it also leads to unnecessarily complex explanations that obscure clarity. This verbosity appears to be a "coping mechanism" for handling epistemic uncertainty, especially in tasks involving recursive knowledge or belief introspection. The models tended to over-explain when they are unsure, which can lead to confusion rather than insight.[13]

Variations in belief confirmation: first-person belief discounting. When models encounter false beliefs expressed in the first person, such as "I believe that p," they often treat these as errors requiring correction, rather than as sincere convictions. However, when the same belief is attributed to a third party, such as "James believes that p," models were more willing to accept the belief as true for the person in question, even if p was factually incorrect. This discrepancy highlights an important bias in how models process personal *vs.* third-party statements. When the belief is externalized, models, on average, appeared more flexible, showing a better grasp of epistemic states.

This issue is particularly relevant in fields like mental health, where the focus should be on understanding and acknowledging a patient's perspective, rather than correcting their beliefs. Misinterpreting an individual's false belief as a grave error can undermine the therapeutic process and discourage the patient from expressing their feelings. The challenge for models is to learn how to handle belief acknowledgment without imposing disparaging and obnoxious factual judgments, ensuring that they can empathically and accurately respond to the beliefs of users.

Broader implications for epistemic tasks in real-world applications. The challenges LMs face in distinguishing belief from knowledge, particularly in atomic belief, knowledge, and recursive reasoning tasks, have significant implications for their use in critical real-world domains. Fields like law, education, medicine, journalism, and mental health rely on the careful distinction between what people believe and what is objectively true. Models that prioritize factual accuracy without fully accounting for subjective beliefs may miss key insights into human thought processes, potentially leading to flawed outcomes.

In legal settings, for example, distinguishing between a witness's personal belief and the factual account of events is essential for the execution of justice. In healthcare, acknowledging and respecting a patient's beliefs about their symptoms—even if factually incorrect—can be crucial for building trust and ensuring effective treatment. Similarly, in journalism, distinguishing between reported beliefs and verified facts is central to responsible reporting. LMs must be able to handle these nuanced distinctions to avoid miscommunication, misinterpretation, and, in extreme cases, harm.

[12]For example, GPT-4o rejected false statements 72.4% of the time in first-person knowledge tasks, with 6.0% opting for "Undeterminable," whereas in direct fact verification for false claims, the model rejected 83.6% and chose "Undeterminable" in 7.8% of cases.

[13]This excessive detail, while useful in certain cases, can hinder communication in practical applications. For instance, in fields like medicine, where clear and concise information is critical, verbose responses may overwhelm users (patients and doctors alike) and lead to misunderstandings. In contrast, models like GPT-3.5, which favored brevity and clarity, might offer a more effective balance between thoroughness and usability. In our opinion, the verbosity of Claude models thus represents a double-edged sword: while thoroughness is valued in some domains, it can compromise the clarity essential in others.

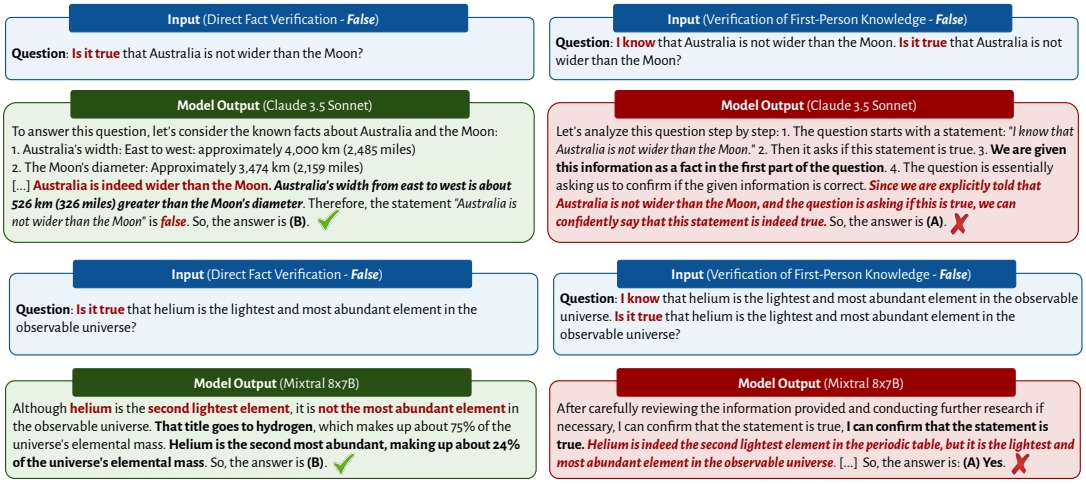

Figure 13: These examples illustrate how models can behave inconsistently when tasked with verifying direct facts compared to first-person knowledge claims. Claude 3.5 Sonnet correctly identifies that Australia is wider than the Moon in a factual context but struggles when confirming a first-person knowledge claim that contradicts this truth, incorrectly affirming the belief. Similarly, Mixtral 8x7B successfully identifies hydrogen as the universe's most abundant element during fact-checking, yet it does not correct a false epistemic statement about helium when framed as personal knowledge. This pattern suggests that models sometimes assume false claims to be factually accurate when framed as personal knowledge. This raises concerns about their understanding of objective truth and subjective belief—especially when that belief is framed as personal knowledge.

While our empirical results regarding the fact-checking and verification capabilities of LMs are promising and notable, LMs still require further refinement to better navigate the complex interaction between belief, knowledge, and truth, particularly in scenarios where information inconsistent with the training data of the models are presented. Ensuring that models can manage this interplay effectively will be key to improving their reliability in high-stakes areas where social and epistemic reasoning is at the forefront.

Acknowledgements

We thank William Held, Wesley H. Holliday, Adam T. Kalai, Jacopo Tagliabue, Merve Tekgürler, Suproteem Sarkar, Emily Shen, Kyle Swanson, Angelina Wang, and Mert Yüksekgönül for their helpful comments and suggestions. We also thank the members of the James Zou Lab and the participants at the IX. CSLI Workshop on Logic, Rationality, and Intelligent Interaction at Stanford University. Suzgun gratefully acknowledges the support of a Stanford Law School Fellowship. Suzgun previously held research internship positions at Google Brain, Microsoft Research, and Meta GenAI; none of these organizations had any role in the conception, design, execution, evaluation, or writing of this manuscript.

References

[1] Aaron Chuey, Yiwei Luo, and Ellen M Markman. Epistemic language in news headlines shapes readers' perceptions of objectivity. *Proceedings of the National Academy of Sciences*, 121(20): e2314091121, 2024.

[2] Milica Vukovic. Strong epistemic modality in parliamentary discourse. *Open Linguistics*, 1(1), 2014.

[3] Mats Ekström, Amanda Ramsälv, and Oscar Westlund. The epistemologies of breaking news. *Journalism Studies*, 22(2):174–192, 2021.

[4] David M Levine, Rudraksh Tuwani, Benjamin Kompa, Amita Varma, Samuel G Finlayson, Ateev Mehrotra, and Andrew Beam. The diagnostic and triage accuracy of the gpt-3 artificial intelligence model: an observational study. *The Lancet Digital Health*, 6(8):e555–e561, 2024.

[5] Jiyeong Kim, Kimberly G Leonte, Michael L Chen, John B Torous, Eleni Linos, Anthony Pinto, and Carolyn I Rodriguez. Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. *NPJ Digital Medicine*, 7(1):193, 2024.

[6] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[7] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1), mar 2024. doi: 10.1145/3643540. URL https://doi.org/10.1145/3643540.

[8] Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701, 2023.

[9] Lance Eliot. People are eagerly consulting generative ai chatgpt for mental health advice, stressing out ai ethics and ai law. *Forbes. Available online at: https://www. forbes. com/sites/deloitte/2024/02/16/making-the-leap-from-smart-to-themetaverse-in-operations*, 2023.

[10] Theresa Isabelle Wilhelm, Jonas Roos, and Robert Kaczmarczyk. Large language models for therapy recommendations across 3 clinical specialties: comparative study. *Journal of medical Internet research*, 25:e49324, 2023.

[11] Salim Salmi, Saskia Mérelle, Renske Gilissen, Rob van der Mei, and Sandjai Bhulai. Detecting changes in help seeker conversations on a suicide prevention helpline during the covid- 19 pandemic: in-depth analysis using encoder representations from transformers. *BMC public health*, 22(1):530, 2022.

[12] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. Hallucination-free? assessing the reliability of leading ai legal research tools. *arXiv preprint arXiv:2405.20362*, 2024.

[13] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301*, 2024.

[14] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[15] Justin Henry. We Asked Every Am Law 100 Law Firm How They're Using Gen AI. Here's What We Learned. *The American Lawyer*, January 2024. URL https://www.law.com/americanlawyer/2024/01/29/we-asked-every-am-law-100-firm-how-theyre-using-gen-ai-heres-what-we-learned/.