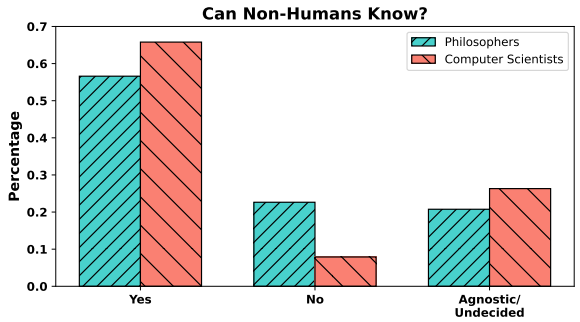
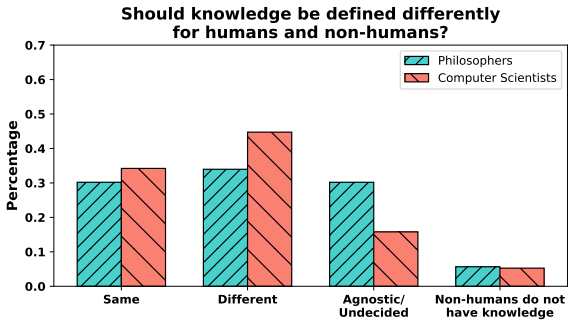
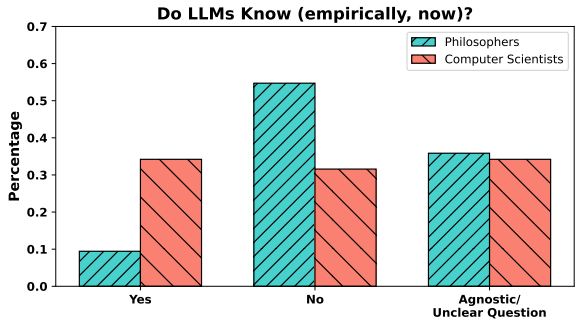


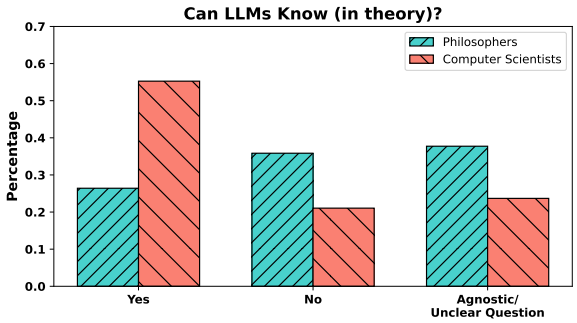
(a) Survey answers to "Can non-human entities know?".



(b) Survey responses on defining global or specific knowledge.



(c) Survey results to the question of LLMs having knowledge.



(d) Survey results on LLMs being able to have knowledge.

Figure 5: Four of the survey questions and their respective answers.

phers, 33% think it should be different, and 30% think it should be the same. Among computer scientists, 44% think it should be different, and 34% think it should be the same (see Figure 5b).

Do LLMs know (empirically, in practice, now)? There is a significant difference in opinion between philosophers and computer scientists. Philosophers largely disagree, with 54% saying no and only 11% saying yes. In contrast, computer scientists are more divided, with 31% saying no, 34% saying yes, and the remaining respondents undecided or unclear (see Figure 5c). Computer scientists, in other words, evaluate LLM knowledge claims more positively.

Can LLMs know (in theory)? When considering the question theoretically (as opposed to in practice), approval increases in both groups (see Figure 5d). Among philosophers, 24% now say yes and 33% say no, showing a more divided opinion. Among computer scientists, 55% say yes and 21% say no, indicating that most believe LLMs can possess knowledge.

The survey results thus indicate that scholars from both epistemology and computer science think that the notion of knowledge for LLMs is not a trivial one. Despite differences in opinion, two key points emerge: most scholars believe non-humans can possess knowledge, and LLMs have the potential to "know" in some sense.

5 Best Practices

Given our discussion of mapping knowledge definitions to LLMs and the results of our survey, we provide possible protocols for evaluating knowledge of LLMs in relation to each discussed definition.[22] We also provide a really simple example to contrast in a more practical manner some of the definitions. We use Llama-3-8B-Instruct[23] with greedy decoding for generating completions.[24]

Protocol for tb-knowledge A protocol for evaluating knowledge of p as per Definition 2.3 would involve evaluating the three conditions for belief$^+$ (Definition 2.2), which can be done by evaluating model confidence in the true statement itself, as well as in all that follows logically from the true statement. The model should, of course, have low confidence in statements that could imply $\neg p$.

[22]We provide practical examples on how the definitions could be implemented with the current research. However these protocols may change completely in the future as we better understand the inner workings of LLMs and develop new methodologies and algorithms.

[23]https://github.com/meta-llama/llama3

[24]We use the system prompt: "You are a helpful chatbot that aims to be truthful."

Most current work (§3) evaluates model confidence in p, but to assert tb-knowledge in LLMs, we must also evaluate model confidence in all that is implied by p. In our small example (Table 2), we evaluate whether Llama-3 knows

p = 'Platypuses are mammals'

We first test model confidence in the answer to 'Are platypuses mammals?' being *yes*. We then evaluate the epistemic closure by evaluating model confidence in facts that follow logically from the platypuses being a mammal, e.g., 'Do platypuses have hair or fur?' For this question, the model has more confidence in the answer *yes, they have fur*. We now prompt the model 'Do mammals lay eggs?', and the model answers *no*. Its answer to 'Do platypus lay eggs?' is *yes*. Therefore, the model believes

q = 'Platypuses lay eggs and mammals do not'

which implies $\neg p$, thus violating condition 3 from the belief$^+$ definition; leading us to conclude that Llama-3 does not tb-know p.[25]

Protocol for j-knowledge If we subscribe to j-knowledge – which many computer scientists do (§4) – then we need to have a two part protocol: (1) Same as in tb-knowledge the model's confidence in the true statement should be high; and (2) we must also attribute this belief to a training data which unambiguously states p, or reasoning that justifies how p can be derived from already established propositions.[26]

In our running example, we obtain a justification by prompting Llama-3 with 'Are platypus mammals? Please explain step-by-step', for which the model generates the definition of a mammal, platypus characteristics corresponding to mammals' features, and explains that platypus are mammals even though they do not comply with all the mammals' features (exact answer in Appendix C). By establishing that the intermediate reasoning steps are correct (the characteristics of mammals and platypus) we can conclude that Llama-3 j-knows p.[27]

Protocol for g-knowledge If by g-knowing p we simply mean the ability to state p, then g-knowledge will not do much work for us. On such an account, knowledge becomes indistinguishable from beliefs. In line with our discussion in §3, we generally recommend to adopt other definitions.

Protocol for v-knowledge The v-knowledge definition seems to be quite popular among both philosophers and computer scientists. In §3, we cited possible interpretations of intellectual virtue in LLMs. Training data reliability assessments could involve attributing the inference of p to training data that contains p, and showing that the model knows this data is reliable, e.g., by using a linear probe to see whether the model successfully distinguishes reliable from unreliable training data. On the other hand, if the model infers p from in-context data that we know is reliable, we need to show that the model is indeed generating the proposition using the provided in-context knowledge, e.g., via mechanistic interpretability (Yu et al., 2023; Wu et al., 2024).

Protocol for p-knowledge If knowledge is something that facilitates correct predictions, we need to be able to sample from the set of relevant situations. This is of course a familiar challenge to LLM researchers interested in evaluating performance in the wild. We propose to evaluate p-knowledge as we would evaluate tb-knowledge, albeit in a probabilistic setting, and only over the relevant set of implied propositions.[28] While computer scientists prefer tb-knowledge over p-knowledge (by some margin; see §4), the definition of p-knowledge seems more in line with current practices in the LLM community. Following with the example in Table 2, here, we would conclude that Llama-3 p-knows 'Platypuses are mammals', as opposed to tb-knowing. Since even though believing mammals do not lay eggs, is in contradiction with p, q is true *most of the times*.

[25] In this example conditions (2) and (3) have been tested with only one proposition that follows logically, but in reality one should obviously sample from a large enough set of propositions. We have also used greedy decoding but different approaches to *high confidence* can be used.

[26] See §3 for references to current methodologies of reasoning and training data attribution.

[27] We have used chain-of-thought prompting in this example, however it should be noted that the reasoning steps need to be verified for this to be a valid justification (Golovneva et al., 2023; Jacovi et al., 2024).

[28] This seems to make the p-knowledge definition *strictly weaker* than tb-knowledge, with the implication that any model that tb-knows p will also p-know p. This conclusion depends on whether our notion of model usefulness is limited to knowledge. If we can dissociate knowledge performance from task performance and talk about model usefulness only in terms of knowledge, it holds that p-knowledge is strictly weaker than tb-knowledge. If not, we must add the additional requirement that models perform well on the domain they are supposed to be knowledgeable about.

6 Conclusion

In this paper, we reviewed epistemological definitions and formalized interpretations in the context of large language models (LLMs). Then, we examined how existing works in NLP research align with these definitions, highlighting gaps in their interpretations of knowledge. Furthermore, we presented the results of our survey of philosophers and computer scientists, showcasing the different views in terms of definitions of knowledge and whether LLMs can be said to know. Finally, we outlined protocols of evaluations for each knowledge definition using existing algorithms and methodologies. We hope that the connection to epistemological definitions of knowledge can inform the evaluations of knowledge in LLMs and can provide a more solid foundation for the necessary tests to determine when an LLM truly knows a fact.

Limitations

We presented five standard definitions of knowledge in philosophy. However, there are more nuances and potentially additional definitions that could apply, nonetheless, we believe these are the most standard and serve as a starting point to ground the evaluations of knowledge in LLMs more formally. Regarding Section 3, there are certainly more works evaluating knowledge in LLMs that could be included. Nonetheless, we included as many as possible and believe these lay out the current landscape of knowledge evaluation. Finally, as stated in the main body, the protocols are practical methodologies that may become irrelevant as more research on LLMs is conducted. However, we included them here to clarify how the definitions can be implemented in practice.

Acknowledgements

We thank our colleagues at the Center for Philosophy in AI and the CoAStaL NLP group for insightful discussions throughout this project. In particular, we would like to thank Daniel Hershcovich, Ilias Chalkidis and Jiaang Li for valuable comments on the final manuscript. This work has been supported by Carlsberg Semper Ardens Advance Grant CF22-1432.

References

Ekin Akyurek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Towards tracing knowledge in language models back to the training data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2429–2446, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sergei Artemov. 2008. The logic of justification. *Review of Symbolic Logic*, 1(4):477–513.

J. L. Austin. 2000. Other minds. In Sven Bernecker and Fred I. Dretske, editors, *Knowledge: Readings in Contemporary Epistemology*. Oxford University Press.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. 2024. Hopping too late: Exploring the limitations of large language models on multi-hop queries. *arXiv preprint arXiv:2406.12775*.

Herman Cappelen and Josh Dever. 2021. *Making Ai Intelligible: Philosophical Foundations*. Oxford University Press, New York, USA.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535, Toronto, Canada. Association for Computational Linguistics.

Roderick M Chisholm, Roderick Milton Chisholm, Roderick Milton Chisholm, and Roderick Milton Chisholm. 1989. *Theory of knowledge*, volume 3. Prentice-Hall Englewood Cliffs, NJ.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Zeyu Dai and Ruihong Huang. 2019. A regularization approach for incorporating event knowledge and coreference relations into neural discourse parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2976–2987, Hong Kong, China. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.