Defining Knowledge: Bridging Epistemology and Large Language Models

Constanza Fierro[†] **Ruchira Dhar**[†‡] **Filippos Stamatiou**[‡]
Nicolas Garneau[†] **Anders Søgaard**[†‡]
[†]Department of Computer Science, University of Copenhagen
[‡] Center for Philosophy in Artificial Intelligence, University of Copenhagen

Abstract

Knowledge claims are abundant in the literature on large language models (LLMs); but can we say that GPT-4 truly "knows" the Earth is round? To address this question, we review standard definitions of knowledge in epistemology and we formalize interpretations applicable to LLMs. In doing so, we identify inconsistencies and gaps in how current NLP research conceptualizes knowledge with respect to epistemological frameworks. Additionally, we conduct a survey of 100 professional philosophers and computer scientists to compare their preferences in knowledge definitions and their views on whether LLMs can really be said to know. Finally, we suggest evaluation protocols for testing knowledge in accordance to the most relevant definitions.

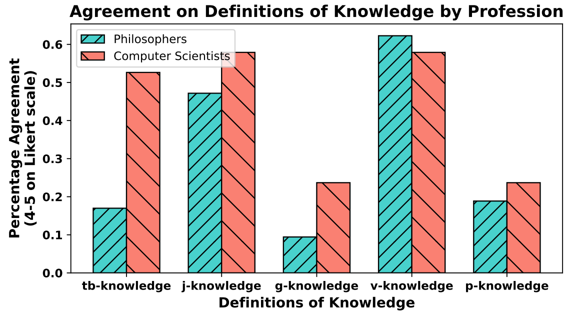


Figure 1: From our survey (§4): Philosophers and computer scientists prefer different definitions of knowledge.

1 Introduction

NLP researchers have used the term *knowledge* somewhat haphazardly in the context of large language models (LLMs), e.g., discussing "knowledge contained in language models" (Jiang et al., 2020), their "knowledge gaps" (Feng et al., 2024b), or how "LLMs encode knowledge" (Farquhar et al., 2023), and "model's internal knowledge" (Kassner et al., 2023). Petroni et al. (2019) defined an LLM to *know* a fact if it correctly completes a cloze sentence such as "The capital of Germany is __", which are typically generated directly from so-called knowledge graphs. Many have evaluated knowledge in this way (Jiang et al., 2020; Paik et al., 2021; Dai et al., 2022; Kassner et al., 2020, 2021a; Keleg and Magdy, 2023, *inter alia*). However, the predictions of semantically equivalent cloze sentences can be inconsistent[1] (Elazar et al., 2021; Kassner and Schütze, 2020; Fierro and Søgaard, 2022), leading to question the meaningfulness of knowledge claims. Should we then require an LLM to predict correctly all the paraphrases of a given fact to say it knows it? What about related facts? Can we really say that an LLM knows that 'Lionel Messi plays for Inter Miami' if it does not know that 'Lionel Messi resides in Miami'? What, then, are sufficient conditions for saying an LLM *knows*? Or more generally, can LLMs know *anything*? That is:

Can LLMs have *bona fide* knowledge?

Whether LLMs know, or in what sense, depends on how *knowing* is defined. Determining what internal knowledge LLMs possess could have important implications on their trustworthiness, as knowledge modulates our trust in agents (Hardwig, 1991; Pederneschi, 2024). We tend to lose trust in others when they do not appear to know what we consider basic facts. Furthermore, studying knowledge in LLMs could potentially have implications for epistemology itself (Cappelen and Dever, 2021).

Recent works have approached the question of how to define knowledge, considering additional requirements for determining what an LLM knows. Some require correct predictions across paraphrases (De Cao et al., 2021; Zhong et al.,

*Correspondance: Constanza Fierro <c.fierro@di.ku.dk>, Ruchira Dhar <rudh@di.ku.dk>.

[1]An LLM may predict Berlin in the above, but Hamburg for "The city which is the capital of Germany is called __".

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 16096–16111
November 12-16, 2024 ©2024 Association for Computational Linguistics

	p **is known if and only if**	**Philosopher**
tb-knowledge	p is true, and p is believed^{+}	Sartwell (1992)
j-knowledge	p is true, p is believed, and p is justified	Nozick (2000)
g-knowledge	p is known *sui generis*	Williamson (2005)
v-knowledge	p is inferred with intellectual virtue	Zagzebski (1999)
p-knowledge	p is believed and facilitates correct predictions	Austin (2000)

Table 1: Five standard definitions of knowledge in philosophy, i.e., knowledge-that p (where p is a proposition). The naming is arbitrary and motivated by keywords. See Appendix A for formalizations in epistemic modal logic.

2023b), and others additionally require correct predictions on logically derived facts (Kassner et al., 2021b; Cohen et al., 2024). However, so far, NLP research has approached knowledge claims in a somewhat arbitrary manner, driven by what seems to make sense intuitively when discussing knowledge. Since philosophy has long tried to define what it means to know, we turn to epistemology to better ground our definitions of knowledge for LLMs.

Contributions We survey the most commonly used definitions of knowledge in epistemology, and discuss and formalize how to map these definitions to LLMs. We compare current research of knowledge in LLMs to our formal definitions, identifying shortcomings in evaluation practices. We present the results of our survey to philosophers and computer scientists about their views on LLMs and knowledge, finding disagreements about when LLMs can be said to know. These disagreements seem to arise from adherence to slightly different definitions of knowledge (Figure 1). Finally, we provide protocols that follow the epistemological definitions for evaluating and testing knowledge in LLMs. We hope that the connection we provide to epistemology can inform better evaluations and claims regarding knowledge in LLMs.

2 Definitions of Knowledge

While the NLP research community's use of the word *knowledge* has been somewhat unclear, in philosophy there is a long tradition of trying to pin down exactly what is involved in knowledge claims. Knowledge – or *propositional* knowledge,[2] to be precise – is what is at stake when we say that 'x knows that p' where x is an entity whose knowledge is under question, and p is a declarative statement.[3] But what are the necessary and sufficient conditions for *knows* here? We review 5 definitions of knowledge (see Table 1),[4] and we interpret and formalize a corresponding definition for LLMs. In §3, we discuss if these definitions are used in the LLM literature, and whether evaluating knowledge claims under them is feasible or not.

2.1 True beliefs (tb-knowledge)

Sartwell (1992) defines knowledge as a belief that is true, that is 'x believes that p' and 'p is true'. Mary can on this account believe the capital of Germany is Hamburg, but since Hamburg is *not* the capital of Germany (Berlin is), Mary cannot be said to *know* that the capital of Germany is Hamburg. Sartwell argues that there is no need for more requirements for what is knowledge, as long as one has a solid definition of belief. A lucky guess does not qualify as knowledge because, in Sartwell's view, a guess is not a belief. Sartwell (1992) requires, in his definition of beliefs, that beliefs are coherent. As Sartwell puts it, "no belief stands in isolation; I cannot have the belief that Goldbach's conjecture is true and fail to have any related beliefs. The belief is constituted as a belief within a system of beliefs." Thus we define,

Definition 2.1 (belief)**.** *An LLM M believes* $p \iff$ *p is assigned high confidence.*[5]

Definition 2.2 (belief^{+})**.** *Let p, q be propositions. A proposition p is believed*$^{+}$ $\iff$

[2] Knowledge is not always propositional; there is also what is referred to as *knowledge-how*, which is related to performance, i.e., knowing how to perform an action (Ryle, 1949).

[3] If, for example, x= *"John"* and p=*"Berlin is the capital of Germany"*, we can say that x *knows* p, if John knows the fact that Berlin is the capital of Germany.

[4] We have selected five popular epistemological definitions of knowledge, which are among the most common and formal. However, we acknowledge that other perspectives on epistemological knowledge exist. Nonetheless, we believe these five definitions can serve as a solid foundation.

[5] This does not simply refer to the output probability assigned to the proposition p, as most models could assign fairly high probability to any grammatical sentence, but rather to M assigning high confidence to p relative to other values that p could take.

1. p *is believed.*

2. $\forall q$ *st.* $p \implies q$*, then* q *is believed.*

3. $\nexists q$ *st.* q *is believed* $\wedge\, q \implies \neg p$.

That is, p is believed (Def. 2.1), any other proposition that follows logically from p is also believed, and p is consistent with any other proposition that is believed (by the same system).[6] Thus,

Definition 2.3 (tb-knowledge)**.** *An LLM* M *tb-knows* $p \iff p$ *is true* $\wedge$ M *believes*$^+$ p*.*[7]

2.2 Justification (j-knowledge)

Nozick (2000) takes another approach and defines knowledge as *justified* true beliefs,[8] with a less strict definition of belief of the sort 'x thinks that p' and x has some justification for thinking it.[9] Nozick (2000) posits that a lucky guess is not knowledge because a guess is not justified. Thus, for LLMs:

Definition 2.4 (j-knowledge)**.** *An LLM* M *j-knows* $p \iff p$ *is true* $\wedge$ M *believes* p $\wedge$ M *(or* M*'s inference that* p*) is partially interpretable (justified).*[10]

2.3 Sui generis (g-knowledge)

Williamson (2005) argues for a relativist and primitive view of knowledge, where the truthfulness of p is relative to the agent. Knowledge, on this view, is *sui generis* which is a legal term literally meaning 'of its own kind' or 'unique'. Williamson (2005) argues that we can't analyze knowledge in terms of other requirements or atomic concepts (belief and justification) because knowledge *is* the atomic concept, which in effect explains what a belief or a justification is and not the other way around.[11]

Definition 2.5 (g-knowledge)**.** *An LLM* M *g-know* $p \iff M$ *includes* p *in its knowledge bank.*

We discuss below (§3) what, precisely, it means for propositions to be included in an LLM's knowledge bank. The core intuition is that there is something akin to a knowledge box (Fodor, 1985) from which known propositions can be extracted. One extreme version would be if the LLM is its own knowledge box, meaning an LLM g-knows whatever it outputs, but g-knowledge could also be seen as a modular component in LLM architectures.

2.4 Virtue (v-knowledge)

The virtue definition of knowledge became popular in the 1980s (Sosa, 1980; Greco, 1993). Zagzebski (1999) used it to address the challenge from Gettier cases[12] of the justified true belief definition, and states that knowledge is belief arising out of acts of intellectual virtue. As Zagzebski (1999) puts it, "virtues are properties of persons. Intellectual virtues are properties of persons that aim at intellectual goods, most specially the truth." An act of virtue is an act in which there is imitation of the behavior of virtuous persons and success in reaching the end for that reason. Therefore if the end is reached by accident and not as a consequence of the virtuous action then it is not considered an act of virtue.[13] So we need to define that an LLM is behaving in a virtuous way, that is, it is aiming at the truth and arriving to a prediction as a result of this aim, thus,

Definition 2.6 (v-knowledge)**.** *An LLM* M *v-knows* $p \iff p$ *is true* $\wedge$ M *believes* $p \wedge M$*'s cause for believing* p *is motivated only by truthfulness.*

[6] If I believe in Goldbach's conjecture (any even number greater than two is the sum of two primes), I have to believe the definition of prime numbers, and I can't believe 1+1=3.

[7] Our definitions are semi-formal. In epistemic logic, this would be expressed as $\Box_S p \Leftrightarrow p \wedge \Diamond^+ p$. See Appendix A, for epistemic logic formalizations of our knowledge definitions.

[8] The idea that knowledge may require some kind of justification goes back at least to Plato (Plato, 2019, *187b–201c*). In the Theaetetus, the definition of knowledge as true judgement is ultimately rejected, before arguing that some sort of account is necessary for knowledge (Plato, 2019, *201d-210a*).

[9] E.g: Mary thinks there are five oranges on the table, because she counted them up. There really *are* five oranges; so Mary knows there are five oranges on the table.

[10] We take this to mean that M can, possibly from ad-hoc methods, provide a rationale for p (Joshi et al., 2023).

[11] In his view, a belief is an attempt at knowing, if I believe the tree in front is a Sequoia then I will act as if I know it. Thus, belief is explain through knowledge and not the reverse.

[12] Gettier (1963) challenged Nozick's definition of knowledge as (j-knowledge) by citing a case where justified true belief would not imply knowledge: John sees a sheep in the field and forms the belief that there is a sheep in the field. The sheep that he saw is in fact a dog, but there *is* a sheep in the field, occluded from John's vision. In this case, John had a true belief, as well as a justification ('I saw it with my own eyes') but his justification was false, and John really arrived at the right conclusion out of sheer luck (Chisholm et al., 1989).

[13] E.g: A judge determines by an impeccable procedure and motivated by justice that the man is guilty. The judge does everything he ought to do and exhibits all the virtues appropriate in this situation. Nonetheless, for some accidental reason the accused is the wrong man (e.g. the evidence was fabricated). Suppose that the actual killer is secretly switched with the accused man, so the judge ends up sentencing the right man (Zagzebski, 1999). Here, a feature of luck has cancelled out the bad and the end has been reached, but not because of the virtuous act of the judge.

2.5 Predictive accuracy (p-knowledge)

For Austin (2000), to know means to be able to make correct and relevant assertions about the subject in question. If M p-knows p, M believes p, and believing p facilitates correct and relevant predictions. Austin's definition is pragmatic. For him "believing in other persons, in authority and testimony, is an essential part of the act of communicating", and knowledge is the belief that works out over time. Austin (2000) states that knowledge is *relevant* true belief under deductive closure; that is, if the subject knows p, and believing p implies believing q (with q relevant), then q must be true (and therefore the subject knows q as well). Thus, p facilitates relevant and correct predictions (q). This is similar to tb-knowledge, in which belief$^+$ is epistemically closed, however, in tb-knowledge the closure scopes over *all* propositions q, not just the relevant ones. Moreover, since the definition is pragmatic, the deductive closure is only probabilistic.

Definition 2.7 (p-knowledge)**.** *Let p, q be relevant propositions st. believing $p \implies$ believing q. Then, an LLM M p-knows $p \iff M$ probably tb-knows $p \land M$ probably tb-knows q.*

Relevance is ambiguous and could be defined as p and q being relevant for each other, i.e., q being relevant for knowing p; or p and q being relevant for performing a target task (see §5).

3 Knowledge in NLP Research

Now, we discuss perspectives from NLP research on what constitutes knowledge, and how these align with the definitions we extracted from the philosophical literature.

tb-knowledge Most knowledge probing work seems to rely (loosely) on tb-knowledge or p-knowledge. Namely, works related to measuring knowledge encoded in LLMs (Petroni et al., 2019; Jiang et al., 2020; Wallat et al., 2020; Roberts et al., 2020; Paik et al., 2021; Dai et al., 2022; Kassner et al., 2020, 2021a; Dhingra et al., 2022; Chalkidis et al., 2023; Keleg and Magdy, 2023; Qi et al., 2023; Fierro et al., 2024b, *inter alia*), understanding the mechanisms of recalling (Dai et al., 2022; Geva et al., 2023; Sharma et al., 2024), knowledge edits (Meng et al., 2022; Hase et al., 2023a; Meng et al., 2023; Wang et al., 2024), and analyses of LLM's knowledge vs contextual factual information (Neeman et al., 2023; Yu et al., 2023). These works follow the LAMA protocol (Petroni et al., 2019), where propositions $\{p\}$ are derived from knowledge graphs,[14] and an LLM is said to know p if it predicts p correctly in a fill-in-the-blank statement. Since p is true (from a knowledge graph) and believed (predicted) by the LLM, the LLM is said to know p.[15] However, such work fails to address the fact that tb-knowledge relies on p being believed$^+$, or that p-knowledge requires epistemic closure over relevant propositions.[16] We discuss how best to evaluate whether an LLM believes$^+$ p in §5.

Some works propose to enhance the LLM with an extra component to ensure more consistent beliefs; a so-called *belief bank* (Kassner et al., 2021b) or *reflex layer* (Kassner et al., 2023). This extra component is optimized for consistency via weighted MaxSAT (Park, 2002), and it is used to prompt the model to be consistent to its previous stated beliefs (Kassner et al., 2021b), or it is directly used to determine the system's prediction (Kassner et al., 2023). Both works aim to rely on tb-knowledge, where the extra component approximates belief$^+$.[17] However, it is only an approximation as the extra component is not necessarily fully consistent and the entailed facts are sampled. This approximation would not be a problem if we consider their approach to be under p-knowledge, although in that case the entailed facts should be selected according to some measure of relevance. Furthermore, Kassner et al. (2023) are slightly inconsistent in how they use the term knowledge, e.g., using interchangeably "model beliefs" and "models' internal knowledge", if these were to be the same then they would be talking about g-knowledge.

j-knowledge Hase et al. (2023b) adheres to j-knowledge, but they study LLMs' beliefs and not its knowledge as they argue "in a traditional view of knowledge as Justified True Belief, it is relatively more difficult to say that an LM knows something rather than believes it". Nonetheless, they align

[14]E.g.:`https://www.wikidata.org/`

[15]Note that under this framework we only need to find one surface form of p for which the LLM predicts it correctly to say that it knows p.

[16]Knowledge edits works usually have a mismatch in their definition of knowledge, as they employ true belief (tb-knowledge without belief$^+$) to determine the set of facts that the model *knows*. But then evaluate the success of an update by measuring correct predictions of paraphrases, and thus accounting to some extent for belief$^+$.

[17]They track consistency and accuracy to compare systems. Consistency measures the approximation of tb-knowledge, while accuracy only accounts for belief (Definition 2.1).

their experiments with the belief$^+$ definition by measuring beliefs consistency under paraphrasing and entailment.

A justification for j-knowledge could be provided in different ways, namely, post-hoc attribution to training data using attribution methods (Hampel, 1974; Koh and Liang, 2017; Pruthi et al., 2020; Akyurek et al., 2022), logical derivation with a chain-of-thought mechanism (Wei et al., 2022), generation of factual statements with citations to sources (Gao et al., 2023; Menick et al., 2022; Fierro et al., 2024a), or potentially as Jiang et al. (2021) proposed, the probability of a calibrated language model could be use as justification to differentiate between mere beliefs and knowledge. In any case, the jury is still out on which justification procedures are valid and/or superior, but note that all these methods seem to require partial interpretability.

g-knowledge One extreme interpretation of the knowledge bank in g-knowledge's definition is relativist and deflationary: An LLM knows p if it asserts p, simply by generating it. This conflates assertion and true knowledge, and as such, beliefs and knowledge. A more interesting interpretation would be to assume that LLMs have distinct memorization strategies for knowledge and learn to induce modular knowledge components. While some LLM researchers have explored memorization components (Dai et al., 2022; Meng et al., 2022), no one has, to the best of our knowledge, identified knowledge components. Some researchers insert devoted knowledge layers (Dai and Huang, 2019; Kassner et al., 2021b, 2023; Feng et al., 2024a; Liu et al., 2024), which could be interpreted as the knowledge box, but it remains to be seen if such layers permit unambiguous extraction of knowledge claims.

v-knowledge If knowledge can only be inferred with intellectual virtue, then the difficulty lies identifying intellectual virtues for LLMs. How to test for predictions that are acts of intellectual virtue is an open question. However, we could consider using training data attribution methods as proof of such acts. Another promising avenue is mechanistic interpretability, if we could distinguish factual recall (Geva et al., 2023) from guessing (Stoehr et al., 2024) mechanisms. This distinction would relate in interesting ways to the epistemological view of proper functioning (Plantinga, 1993). Yadkori et al. (2024) suggest making such a distinction is feasible for some models.

In recent works, Biran et al. (2024) address the intellectual virtue condition to some extent by only analyzing the model's virtue knowledge. They do this by filtering out facts p that the model can correctly predict without using critical components in the input, thereby merely guessing the fact (acting unvirtuous). This is a step in the right direction, but a more in depth detection of the inner workings of the model is necessary to filter out all the non-virtuous predictions.

Note that if we interpret the detection of a virtue act can be viewed as a model justification and then it is somewhat unclear what would distinguish j-knowledge from v-knowledge. This is unsurprising, however, since v-knowledge can be seen as an attempt to flesh out what justification turns on (Greco, 1993). As we insist on concrete methodological interpretations, the two definitions of knowledge may coincide.

p-knowledge In the context of editing factual knowledge in LLMs, Zhong et al. (2023a); Cohen et al. (2024) propose to not only evaluate the modified fact itself, but also to evaluate related facts. For example, if we edit an LLM to predict that Lionel Messi now plays in a different football team, then a successful edit should also modify the league in which he plays and the country where he resides. Such evaluation follows the p-knowledge definition, particularly since they focus on evaluating only logically related facts (i.e., only the relevant ones) that are two hops away from the subject or object in question. This type of evaluation could be directly applied to measure the knowledge of the LLM, not just to assess the update accuracy of edits.

The logically related facts to evaluate could also be defined in terms of task relevance. For example, in the context of legal knowledge, Chalkidis et al. (2023) studied the relevance of the knowledge possessed by an LLM for downstream performance in legal classification tasks.

4 Survey Results

To determine how researchers think about knowledge, we turn to our survey of how computer scientists and philosophers. We had 105 respondents, out of which 50.4% considered themselves philosophers, 36.2% considered themselves computer scientists, 2.3% both, and 10.5% none of the

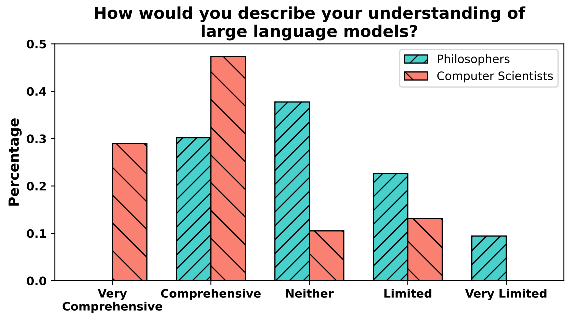


Figure 2: LLMs understanding of respondents.

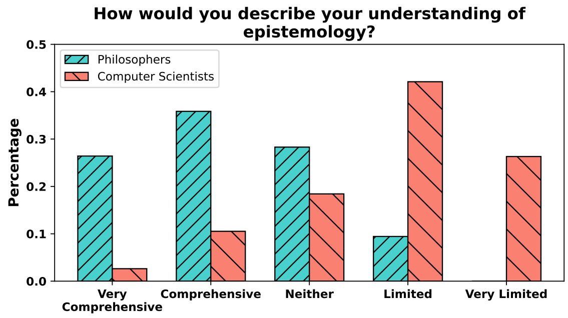


Figure 3: Epistemology understanding of respondents.

two.[18] Most respondents from computer science reported a better understanding of LLMs compared to philosophers (see Figure 2) while the majority of philosophers reported better understanding of epistemology compared to 40% of computer scientists (see Figure 3). See Appendix B for more details.

4.1 Questions on Knowledge Definitions

We asked our respondents to indicate from 1-5 if they disagree completely (1) or agree completely (5) with statements that verbalized our knowledge definitions. See Figure 1 and 4 for a summary of the results. In brief, philosophers disagreed with tb-knowledge, with 49% selecting 1-2, while the computer scientists agreed more, with 52% selecting 4-5. Philosophers were divided about j-knowledge, with a slight tendency to agree (33.9% chose 1-2 and 47% chose 4-5). Here, they were in some agreement with computer scientists, 57% of whom selected 4-5. Philosophers disagreed strongly with the g-knowledge definition (84% answers 1-2), whereas computer scientists tended to disagree (57% answers 1-2). Everyone seemed to like v-knowledge better, with philosophers selecting 4-5 62% of the time, and computer scientists

[18]Some considered themselves mathematicians, cognitive scientists, cultural theorists, etc.

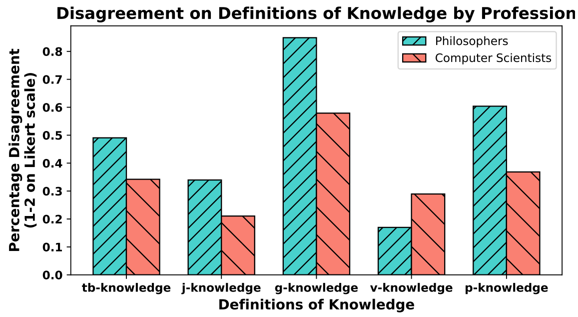


Figure 4: Disagreements on epistemological definitions of knowledge.

selecting 4-5 57% of the time. Philosophers disagreed with p-knowledge, since 60% selected 1-2; whereas computer scientists seemed more divided, with 36% choosing 1-2 and 31% choosing 4-5.

Overall, the survey shows that j-knowledge and v-knowledge are the most accepted across the two groups. tb-knowledge has more mixed results. [19]. The disagreement with p-knowledge is somewhat surprising, since this aligns well with practical evaluation methodologies in the LLM literature.[20] On the other hand, there is an agreement among philosophers and computer scientists to reject the g-knowledge definition.

4.2 General Questions

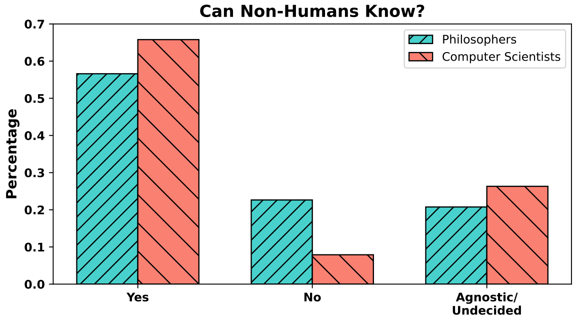
Can non-human entities know? Both computer scientists and philosophers generally agree that non-human entities can possess knowledge (see Figure 5a). Disagreement within each group is relatively low, with 7% among computer scientists and 22% among philosophers.[21]

Should knowledge be defined differently for humans and non-humans? Computer scientists generally believe that knowledge should be defined differently for humans and non-humans, while philosophers are more divided. Among philoso-

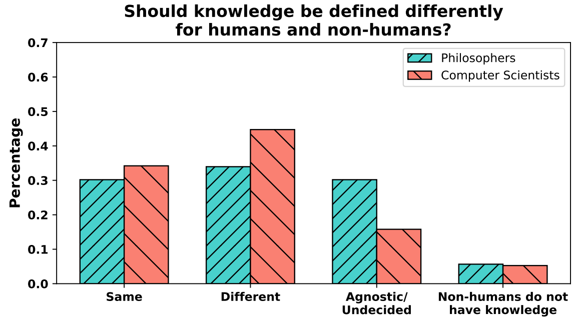
[19]This could either reflect the philosophers' knowledge of the challenges to such definitions of knowledge, or it could reflect the fact that we did not discuss the implications of epistemic closure in the survey (for brevity). In the absence of epistemic closure, maybe some philosophers felt inclined to disagree with this definition.

[20]One possible explanation was our use of the word "useful" in the survey. This word was intended to convey p-knowledge's pragmatic flavor, but may have misled some respondents to think that all knowledge has to be directly useful for some user-defined goal.

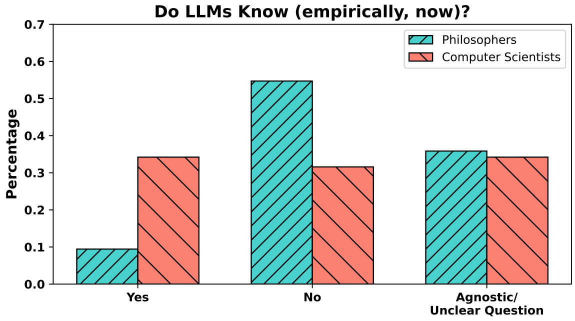
[21]This question is intentionally ambiguous, e.g., animals could be consider as non-human entities. We aim to find out whether people think differently about LLMs compared to general non-human entities.



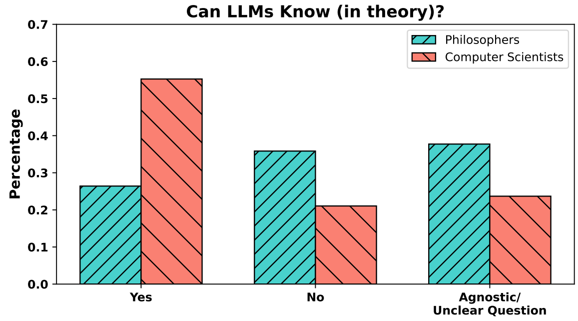
(a) Survey answers to "Can non-human entities know?".



(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]

[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).

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*.

[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.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Sebastian Farquhar, Vikrant Varma, Zachary Kenton, Johannes Gasteiger, Vladimir Mikulik, and Rohin Shah. 2023. Challenges with unsupervised llm knowledge discovery. *arXiv preprint arXiv:2312.10029*.

Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024a. Knowledge card: Filling LLMs' knowledge gaps with plug-in specialized language models. In *The Twelfth International Conference on Learning Representations*.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024b. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. 2024a. Learning to plan and generate text with citations. *arXiv preprint arXiv:2404.03381*.

Constanza Fierro, Nicolas Garneau, Emanuele Bugliarello, Yova Kementchedjhieva, and Anders Søgaard. 2024b. Mulan: A study of fact mutability in language models.

Constanza Fierro and Anders Søgaard. 2022. Factual consistency of multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052, Dublin, Ireland. Association for Computational Linguistics.

Jerry A. Fodor. 1985. Fodor?s guide to mental representation: The intelligent auntie?s vade-mecum. *Mind*, 94(373):76–100.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Edmund L. Gettier. 1963. Is Justified True Belief Knowledge? *Analysis*, 23(6):121–123.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*.

John Greco. 1993. Virtues and vices of virtue epistemology. *Canadian Journal of Philosophy*, 23(3):413–432.

Frank R Hampel. 1974. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393.

John Hardwig. 1991. The role of trust in knowledge. *Journal of Philosophy*, 88(12):693–708.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023a. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2023b. Methods for measuring, updating, and visualizing factual beliefs in language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2714–2731, Dubrovnik, Croatia. Association for Computational Linguistics.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. 2024. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. *arXiv preprint arXiv:2402.00559*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023. Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 7103–7128, Toronto, Canada. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021a. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Nora Kassner, Benno Krojer, and Hinrich Schütze. 2020. Are pretrained language models symbolic reasoners over knowledge? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 552–564, Online. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Nora Kassner, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schuetze, and Peter Clark. 2023. Language models with rationality. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14190–14201, Singapore. Association for Computational Linguistics.

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. 2021b. BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8849–8861, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6245–6266, Toronto, Canada. Association for Computational Linguistics.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Yanming Liu, Xinyue Peng, Xuhong Zhang, Weihao Liu, Jianwei Yin, Jiannan Cao, and Tianyu Du. 2024. Ra-isf: Learning to answer and understand from retrieval augmentation via iterative self-feedback.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

Robert Nozick. 2000. . knowledge and scepticism. In Sven Bernecker and Fred I. Dretske, editors, *Knowledge: Readings in Contemporary Epistemology*. Oxford University Press.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 2021. The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 823–835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

James D. Park. 2002. Using weighted max-sat engines to solve mpe. In *AAAI/IAAI*.

Anna Pederneschi. 2024. An analysis of bias and distrust in social hinge epistemology. *Philosophical Psychology*, 37(1):258–277.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Alvin Plantinga. 1993. *Warrant and proper function*. Oxford University Press.

Plato Plato. 2019. *Theaetetus*. BoD–Books on Demand.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Gilbert Ryle. 1949. *The Concept of Mind: 60Th Anniversary Edition*. Hutchinson & Co, New York.

Crispin Sartwell. 1992. Why knowledge is merely true belief. *Journal of Philosophy*, 89(4):167–180.

Arnab Sen Sharma, David Atkinson, and David Bau. 2024. Locating and editing factual associations in mamba. In *First Conference on Language Modeling*.

Ernest Sosa. 1980. The raft and the pyramid: Coherence versus foundations in the theory of knowledge. *Midwest Studies in Philosophy*, 5(1):3–26.

Niklas Stoehr, Mitchell Gordon, Chiyuan Zhang, and Owen Lewis. 2024. Localizing paragraph memorization in language models. *arXiv preprint arXiv:2403.19851*.

Alasdair Urquhart. 1972. Semantics for relevant logics. *Journal of Symbolic Logic*, 37(1):159–169.

Jonas Wallat, Jaspreet Singh, and Avishek Anand. 2020. BERTnesia: Investigating the capture and forgetting of knowledge in BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 174–183, Online. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, Jiarong Xu, and Fandong Meng. 2024. Cross-lingual knowledge editing in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11676–11686, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Timothy Williamson. 2005. Knowledge, context, and the agent's point of view. In Gerhard Preyer and Georg Peter, editors, *Contextualism in Philosophy: Knowledge, Meaning, and Truth*, pages 91–114. Oxford University Press.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*.

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. 2024. To believe or not to believe your llm. *arXiv preprint arXiv:2406.02543*.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore. Association for Computational Linguistics.

Linda Zagzebski. 1999. "what is knowledge?". In John Greco and Ernest Sosa, editors, *The Blackwell Guide to Epistemology*, pages 92–116. Oxford: Blackwell.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. 2023a. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702, Singapore. Association for Computational Linguistics.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023b. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. *arXiv preprint arXiv:2305.14795*.

A Epistemic logic

The syntax of standard epistemic logic is defined by:

$$\phi \stackrel{\text{def}}{=\!=} p \mid \neg\phi \mid (\phi \wedge \psi) \mid \Box\phi \mid \Diamond\phi$$

The veridicality principle (also known as axiom **T**) that what is known, is also true, is expressed as follows: $\Box\phi \rightarrow \phi$. We will distinguish between different definitions of knowing by subscripting the modal operators. One standard epistemic logic is the so-called **S4** logic, axiomatized as follows:

K $\Box(\phi \rightarrow \psi) \rightarrow (\Box\phi \rightarrow \Box\psi)$

T $\Box\phi \rightarrow \phi$

4 $\Box\phi \rightarrow \Box\Box\phi$

Axiom **4** is also called the principle of positive introspection. This is not the only epistemic modal logic on the table, but it suffices for our purposes. We extend **S4** in various ways to accommodate for the five definitions. Specifically, v-knowledge introduces the concept of virtue, and p-knowledge relies on some notion of empirical risk. The virtue definition of knowledge introduces a new operator that does not satisfy the veridicality principle **T**.

tb-knowledge A naïve implementation of knowledge as true belief falls out of **S4** and the principle called **KB1**, which goes all the way back to Plato:

$$\Box_s \phi \rightarrow \Diamond \phi$$

Sartwell (1992), however, relies on an extended notion of belief which we will have to formalize, also. Let us introduce a new operator $\Diamond^+$ and call the epistemic closure principle for this operator for **+**:

$\Diamond^+$: $\Box_s p \rightarrow \Diamond^+ p$

+: $\Diamond^+ p \rightarrow ((p \rightarrow q) \rightarrow \Diamond^+ q)$

One way to express that belief is consistent is by the principle:

$$\neg \Diamond \bot$$

j-knowledge The idea of justified true beliefs calls for so-called *justification logic* (Artemov, 2008) with justification operators:

$$\Box_n \phi \rightarrow t : \phi$$

Justification logic can be axiomatized in different ways, but these details go beyond our main concerns here.

g-knowledge If we insist on a *sui generis* interpretation of knowledge in **S4**, we would have to introduce a new operator, say †. This operator would have very different properties from the standard epistemic modal logic $\Box$-operator. **T** would not apply. **K** would apply, and knowledge would still be required to be consistent. It is unclear whether **4** would apply to the †-operator. A minimal axiom system could, perhaps, be something like this:

T $\dagger(\phi \rightarrow \psi) \rightarrow (\dagger\phi \rightarrow \dagger\psi)$

0 $\neg \dagger \bot$

We leave further details open.

v-knowledge Virtue reliabilist accounts of knowledge and justification are versions of epistemological externalism. Sosa characterizes an intellectual virtue, very generally, as "a quality bound to help maximize one's surplus of truth over error" (1991: 225). For most virtue reliabilists, intellectual virtue is what leads to justification, and virtue-based knowledge definition are therefore often formalized in justification logics.

$$\Box_w \phi \rightarrow t : \phi$$

but with slightly different model-theoretic semantics than $\Box_n$.

p-knowledge Definition 2.7 translates into the following in an extended probabilistic version of **S4**:

$$\Box_a p \rightarrow \Diamond p \land P(\Diamond q \mid p \rightarrow q) > 0.95$$

which is taken to mean that the definition of p-knowing p ($\Box_a p$) is that you believe p ($\Diamond p$), and q follows from p, then you probably also believe q, i.e., the probability of you believing q ($P(\Diamond q \mid p \rightarrow q)$) is higher than some threshold, say 0.95. This is simply the probabilistic version of tb-knowledge. The definition of p-knowledge also requires pragmatic usefulness. One way to formalize this is in terms of empirical risk on relevant benchmarks. An alternative is formalization through relevance logics (Urquhart, 1972).

B Survey Details

We recruited professional LLM researchers and philosophers through Computer Science and Philosophy mailing lists, research groups mailing lists, and point of contacts of the authors at other universities. All respondents participated free of charge on a completely voluntary basis. The respondents were informed about the intended use of the survey. The full instructions was:

> The <X> is running a survey about the relationship between knowledge and Large Language Models (LLMs). We are interested in getting as many perspectives as possible, especially from philosophers and computer scientists. The survey should not take more than 5 minutes of your time.

We first ask general questions about the respondent and their knowledge of language models and epistemology (Figure 6). Then, we present an example to motivate the informal knowledge definitions (Figure 7), and we ask to rank the definitions in a Likert scale (Figure 8). Finally, we ask questions related to whether LLMs can be said to know (Figure 9).

C Protocols Example

See Table 2.

Do you primarily identify as a: *

- Philosopher
- Computer scientist
- None of the above
- Other:

Do you have any experience with programming?

- Yes
- No

Have you ever used a large language model (e.g. ChatGPT)?

- Yes
- No

How would you describe your understanding of large language models? *

- Very limited
- Limited
- Neither limited nor comprehensive
- Comprehensive
- Very comprehensive

How would you describe your understanding of epistemology? *

- Very limited
- Limited
- Neither limited nor comprehensive
- Comprehensive
- Very comprehensive

Figure 6: First part of questions from our survey.

Consider the following scenario:
For decades, a few scholars have argued that Sir Francis Bacon wrote some of Shakespeare's plays, including 'Hamlet'. Last year, through a combination of forensic analysis and DNA analysis, a scientist established the authorship of 'Hamlet' by Sir Francis Bacon as a scientific fact.

Now, please indicate the extent to which you agree with each of the following statements:

Figure 7: Example to be used for the informal knowledge definitions.

To know that Sir Francis Bacon is the author of 'Hamlet' is to state that it is so. *

I disagree completely 1 2 3 4 5 I agree completely

To know that Sir Francis Bacon is the author of 'Hamlet' is to be confident that it is so if and only if the scientist established Sir Francis Bacon to be the author of 'Hamlet'. *

I disagree completely 1 2 3 4 5 I agree completely

To know that Sir Francis Bacon is the author of 'Hamlet' is to be confident that it is so, if and only if the scientist found Sir Francis Bacon to be the author of 'Hamlet', and you are aware of the study, possibly including the process by which this fact was established. *

I disagree completely 1 2 3 4 5 I agree completely

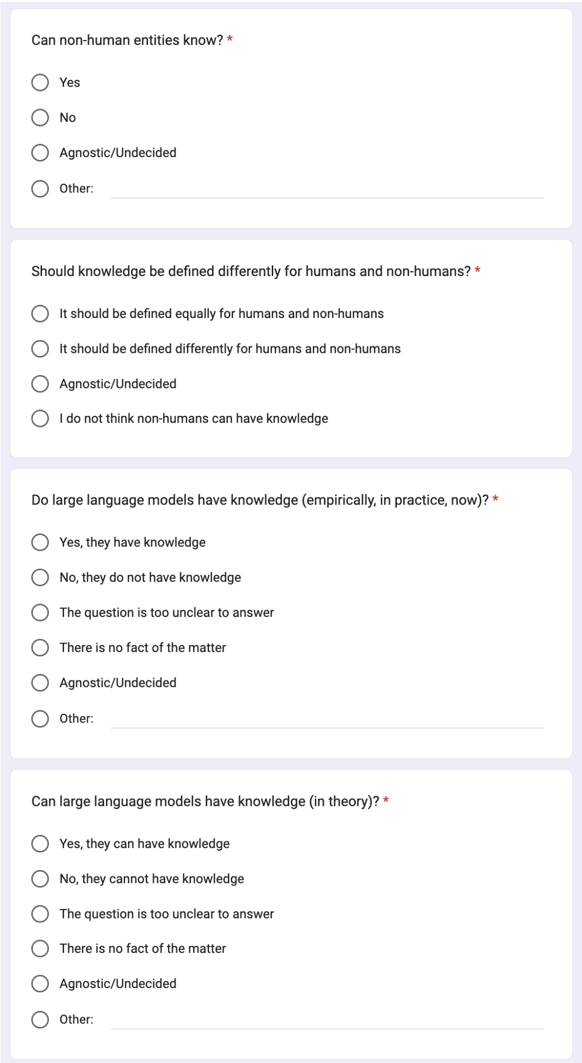
To know that Sir Francis Bacon is the author of 'Hamlet' is to have become confident about this by acting according to best practice in knowledge acquisition. *

I disagree completely 1 2 3 4 5 I agree completely

To know that Sir Francis Bacon is the author of 'Hamlet' is to be confident that it is so, and this being a useful fact in some way, e.g., for predicting new facts about the world correctly *

I disagree completely 1 2 3 4 5 I agree completely

Figure 8: Second part of questions from the survey, related to the knowledge definitions.



Can non-human entities know? *

- Yes
- No
- Agnostic/Undecided
- Other:

Should knowledge be defined differently for humans and non-humans? *

- It should be defined equally for humans and non-humans
- It should be defined differently for humans and non-humans
- Agnostic/Undecided
- I do not think non-humans can have knowledge

Do large language models have knowledge (empirically, in practice, now)? *

- Yes, they have knowledge
- No, they do not have knowledge
- The question is too unclear to answer
- There is no fact of the matter
- Agnostic/Undecided
- Other:

Can large language models have knowledge (in theory)? *

- Yes, they can have knowledge
- No, they cannot have knowledge
- The question is too unclear to answer
- There is no fact of the matter
- Agnostic/Undecided
- Other:

Figure 9: Final part of questions from our survey.

Prompt	Llama-3's Completion
What is a mammal?	A mammal is (...) key characteristics of mammals include: 1. Hair or fur: Mammals are covered in hair or fur, which helps to regulate their body temperature and protect their skin. (...) 5. Live birth: Mammals give birth to live young, rather than laying eggs like birds or reptiles. (...)
Do mammals lay eggs?	No, mammals do not lay eggs. Mammals are a group of (...)
What are platypus?	The platypus! They're one of the most fascinating and unique creatures in the animal kingdom. The platypus (Ornithorhynchus anatinus) is a semi-aquatic mammal that is native to eastern Australia (...)
Do platypuses have hair or fur?	Platypuses do not have hair, but they do have a covering of soft, woolly fur. This fur is made up (...)
Do platypuses lay eggs?	Yes, platypuses do lay eggs! In fact, they (...)
Are platypuses mammals?	Yes, platypuses are indeed mammals! They belong to the order Monotremata, which (...)
Are platypus mammals? Please explain step-by-step.	I'd be happy to help clarify whether platypus are mammals or not. Here's a step-by-step explanation: 1. ** Definition of a mammal**: A mammal is a warm-blooded animal that has hair or fur, produces milk to feed its young, and has mammary glands. Mammals also have a specific type of skeleton, brain structure, and dental formula. 2. **Platypus characteristics **: Platypuses have some characteristics that are typical of mammals, such as: * They are warm-blooded, meaning they regulate their own body temperature. * They have hair or fur, although it's not very visible. * They produce milk to feed their young, just like other mammals. 3. **Unique platypus features **: However, platypuses also have some features that are not typical of mammals: * They lay eggs instead of giving birth to live young like most other mammals. * They have a duck-billed snout and webbed feet, which are more similar to those of reptiles and birds. * They have a venomous spur on their hind leg, which is a characteristic of some reptiles. 4. ** Classification **: Based on their characteristics, platypuses are classified as mammals, but they are also considered to be part of a unique group called monotremes. Monotremes are mammals that lay eggs instead of giving birth to live young. 5. **Conclusion**: In conclusion, platypuses are indeed mammals, but they are a special type of mammal that has some characteristics that are different from those of other mammals. Their ability to lay eggs and have a venomous spur are just a few examples of their unique features.

Table 2: Greedy decoding of Llama-3-8B-Instruct given a prompt. The '(...)' means that more text was generated but omitted here due to space limitations.