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.

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:

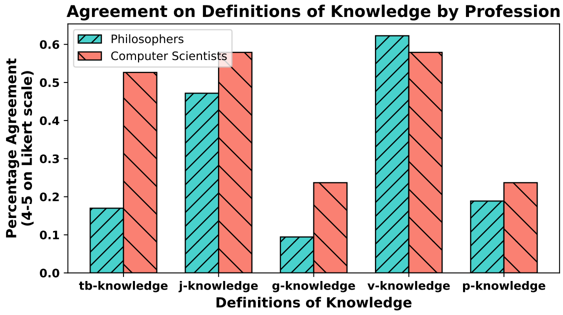

Figure 1: From our survey (§4): Philosophers and computer scientists prefer different definitions of knowledge.

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.