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 $p \wedge M$ 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

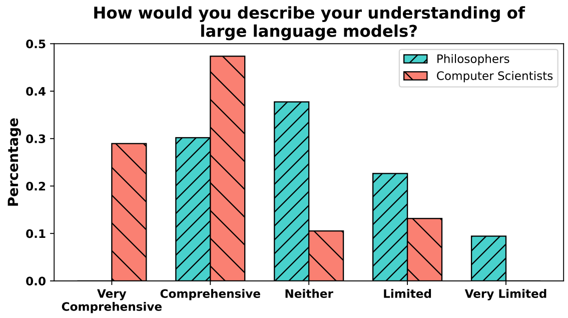

Figure 2: LLMs understanding of respondents.

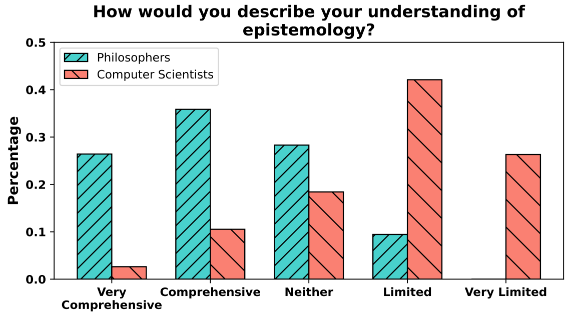

Figure 3: Epistemology understanding of respondents.

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 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.

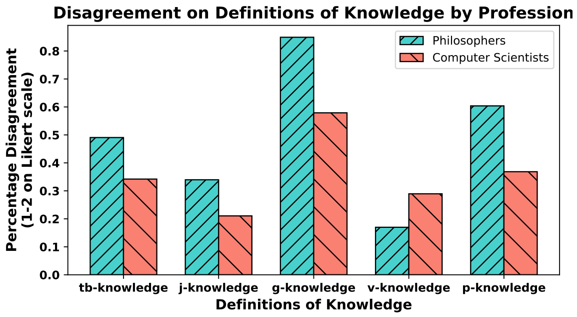

Figure 4: Disagreements on epistemological definitions of knowledge.

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-

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

[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.