Can Large Language Models Know What They Know? Sosa, Aptness, and Mechanistic Evidence

Jonas Strabel

PTH Sankt Georgen

August 4, 2026

Keywords: mechanistic interpretability · large language models · philosophy of mind

1 Introduction

At least since Plato, the great value of knowledge has been recognized [1, 97a–98a]. It is the character of knowledge, as opposed to mere true belief, that allows us to *depend* on it.

With artificial intelligence rapidly becoming ubiquitous [2], it is time to ask whether artificial systems can be said to have knowledge [3]. Answering this question has direct implications for the amount of trust we can place in these systems.

However, what counts as knowledge remains contested to this day [5, ch. 2.3]. This paper therefore adopts one recent and influential conception of knowledge: Ernest Sosa's virtue epistemology [6].

2 Sosa's Framework of Knowledge

Sosa analyzes performances with an aim, such as an archer's shot at a target. Performances can be assessed along the AAA structure: "accuracy: reaching the aim; adroitness: manifesting skill or competence; and aptness: reaching the aim *through* the adroitness manifest" [6, p. 23]. Performances need not be intentional: a heartbeat succeeds if it pumps blood, though the heart presumably does not intend to do so [6, p. 23].[1] Performances comprise not only acts but also sustained states, including beliefs [6, p. 24].

For the present comparison, I will use the following working operationalization of aptness.

[1] This removes a potential hurdle: it has been argued that intention is necessary for meaningful speech or understanding [7], but since Sosa's notion of performance does not require intention, that debate need not be settled here.

A performance P by a subject A is apt only if: (C1) A possesses a competence C, a disposition to succeed in a relevant range of appropriately normal conditions; (C2) P succeeds; (C3) the present conditions fall within that appropriate range; and (C4) P succeeds through the manifestation of C.

Animal knowledge K is the special case of apt performance in the way of belief: P is a belief, and its accuracy is its truth.

Writing K for animal and K^+ for reflective knowledge, the basic idea is $K^+p \leftrightarrow KKp$ [6, p. 32], though Sosa offers this as a merely first approximation; the full account admits of degrees.

Reflective knowledge adds an apt higher-order belief B_M that the first-order belief B_p is apt. Applying the same structure at the second order gives a working test: (R1) B_p is apt; (R2) A has a meta-competence for assessing the aptness of beliefs of the relevant kind; (R3) B_M is accurate (i.e. B_p really is apt) under conditions appropriate for that meta-competence; and (R4) the accuracy of B_M manifests that meta-competence. The relevant content is aptness, not merely the truth of p.

3 Belief

A framework on which knowledge is defined as apt *belief* obviously hinges on the attribution of belief, so let us address that question first. Whether LLMs have beliefs is highly controversial [3, pp. 2, 8] and has been discussed extensively in the literature. Since belief is not the main subject of this paper, I will examine only a recent survey, two papers that directly address the question, and some positions that are particularly useful for present purposes. The aim is not to settle the debate about belief, but to determine whether it is plausible to assume, for the purposes of this paper, that LLMs[2] can have beliefs.

Herrmann and Levinstein propose four criteria for belief-like representations in LLMs: accuracy, coherence, uniformity, and use [8]. The same authors confront the more fundamental question of whether LLMs can have beliefs at all in [9, Ch. 5]. Against skeptical arguments in the tradition of Bender et al. and Shanahan, they reject the inference from next-token-prediction training to an inability to track truth or form beliefs: tracking features of the world can itself be useful for predicting text, though the strength of this pressure varies across contexts. Their conclusion is that the presence of beliefs is an empirical question current methods do not yet settle reliably [9, Ch. 5].

[2] I do not add another general introduction to large language models. The term is by now ubiquitous, and I discuss the relevant technical details where they become necessary for the analysis.

At the opposite end of the debate, Cappelen and Dever defend a deliberately strong position: starting from ordinary descriptions of LLM behaviour such as answering questions, their holistic network premise holds that the mental capacities these descriptions presuppose must be attributed together – if an LLM genuinely answers a question, it knows the answer, and if it knows, it believes [4].

The recent survey by Millière and Buckner examines these views within a wider spectrum of eliminativist, fictionalist, interpretivist, and realist positions. Their discussion of propositional attitudes again leaves the issue unsettled [3, Ch. 7]. Goldstein and Levinstein reach a similarly cautious conclusion. They argue that LLMs possess robust internal representations under several prominent philosophical accounts of representation, while leaving open whether these representations belong to a stable belief–desire–action psychology [12]. The evidence therefore supports a representational attribution more strongly than it supports the attribution of complete folk-psychological belief.

Chalmers offers a useful middle path here: "generalized propositional attitudes" that abstract from the specifically human connotations of belief, letting us discuss LLMs' causally efficacious internal representations without first settling whether they have minds in the human sense [13, p. 21].

Sosa's own later work offers a possible bridge from these minimal representations to belief. In *Judgment and Agency*, he distinguishes judgmental belief from "merely functional" representation [14, pp. 80, 92–93]. Merely functional representations are implicit, action-guiding, and aimed at correctness through their functional role, without requiring direct voluntary control. This is useful for this paper because, even if the question of human-like beliefs in LLMs is settled negatively in the future, Sosa's lowest category of belief does not require them. The distinction is therefore compatible with the 2007 framework adopted in this paper: a merely functional belief may in principle constitute animal knowledge, while reflective knowledge remains a further question.

The available literature thus does not establish that current LLMs have beliefs in the full human sense. Nevertheless, the attribution of at least some belief-like, action-guiding representations is neither arbitrary nor implausible. I will therefore adopt as a working hypothesis that some internal states of LLMs can qualify as beliefs, at least in Sosa's minimal functional sense. The conclusions that follow are conditional on this hypothesis: if LLMs lack even functional beliefs, they cannot possess knowledge on Sosa's account. If they have them, we can ask whether these beliefs are apt and hence are animal knowledge, and whether any are also aptly recognized so as to amount to reflective knowledge.

4 Animal Knowledge in LLMs

Apt performance as such sets a low bar. Even a pocket calculator performs aptly when it calculates correctly: it succeeds, it has a stable disposition to succeed, and its successes manifest that disposition. The reason it has no knowledge in the epistemic sense is that knowledge is apt performance *in the way of belief*, and a calculator has no beliefs. That LLMs perform aptly in *some* way is as trivially true as in the calculator case. The interesting question is whether they perform aptly in the way of belief.

What separates epistemic from non-epistemic performances? Human performances can contain epistemic components: an archer need not consciously reason, yet her shot can be guided by a belief about where the arrow should go. Suitable epistemic attributions can be found inside performances of many kinds.

Since Section 3 granted belief-apt states to LLMs as a working hypothesis, one might suspect that the question of animal knowledge has now become trivial. Two remarks on that suspicion. First, even if the answer were nearly trivial *within Sosa's framework*, that would itself be a finding. Under the classic justified-true-belief conception, nothing about reliable success settles whether a system possesses justification, and the question of LLM knowledge would remain wide open. That an externalist virtue epistemology is this permissive toward reliable artificial performers seems to be a fact about externalism. Second, the answer is not in fact trivial, because a performance's truth must be attained *through* a competence (C4) whose appropriate range covers the present case (C3). The interesting action is therefore in C1, C3, and C4. I take the conditions in turn.

4.1 C1: Competence

LLM training happens in two large phases. In pre-training, the model is tasked with predicting enormous amounts of text; where it errs, its weights are adjusted to fit the text better and be more accurate next time [15]. By analogy, this resembles a child learning to speak: with each correction, speech becomes more accurate, and thought and language develop together. This phase is by far the largest by volume and gives the model both its command of language and most of its factual knowledge.

In post-training, specific behaviour, response style, and alignment are trained: human demonstrations and preference feedback teach the model how to react to the tasks users actually give it [15]. The analogy here is on-the-job training: the graduate has already acquired language and knowledge; now she learns to do her job well.[3]

[3]One might worry that since the polished capability only becomes visible after the comparatively small post-training phase (post-trained models are vastly preferred by users over their pre-trained bases [16]), the competence is really an artifact of that phase. But pre-training builds the competence and post-training adapts it for specific use, much as on-the-job training turns a graduate's existing abilities into strong performance on the job.

Both phases have been refined and scaled up greatly in recent years, and the resulting language capability, knowledge, and task performance of frontier models are vast [2].

Sosa's competence is a *disposition* seated in the agent, and the trained weights are exactly that: a standing structure that, given an input, makes certain successes highly likely. Just as a human at any moment is aware of only a fraction of what she knows, a model has no occurrent activity between inferences; the competence persists in the weights and manifests during inference, that is, during the model's "performance."

4.2 C2: Accuracy

Accuracy in the way of belief is truth, and it is assessed per belief. Most people agree that humans can have knowledge, yet humans are not always accurate; the same goes for LLMs. Benchmark results document that frontier models answer very wide ranges of questions correctly [2, #1], which matters here only as evidence that accurate performances exist in abundance. Whether any particular accurate performance amounts to knowledge is settled by the remaining conditions.

4.3 C3: Appropriateness

Whether conditions are appropriately normal is relative to the competence. Modern LLMs are trained on enormous amounts of text, facts, and real-world tasks, so we should expect many ordinary situations to fall within the range of some competence.

The competence-relativity of appropriateness also explains the seemingly paradoxical capability profile, discussed as a variant of Moravec's paradox or, for LLMs, as the "jagged frontier": models outperform nearly all humans in domains traditionally taken to require the most intelligence, such as mathematics and programming [2], while losing to attentive laypeople on trick questions requiring everyday physical and social intuition [17]. If appropriateness were indexed to human difficulty, this would be paradoxical. Indexed to the model's competences, it is what one should expect: models are good at what they are trained at and at the tasks their architecture enables them to perform well.

Clearly excluded are questions about private matters – anything absent from training data and available context. Asking a model what happened yesterday in one's house is a question it cannot answer, but note the correct diagnosis: it is not that such a question puts a general answering competence into "inappropriate conditions"; rather, no competence of the model has such facts in its range at all, so there is nothing whose manifestation could make the answer likely true. What a model *can* have is a second-order disposition to flag such questions, and current models are trained, with some success, to do exactly that [18, p. 13]; I return to this in Section 5.

Conditions can also be made inappropriate from the inside. Interpretability researchers can amplify or suppress individual features of a model, a practice roughly comparable to a human taking a substance, though far more targeted. When Templeton et al. strongly activated the feature representing the Golden Gate Bridge, the resulting "Golden Gate Claude" wove the bridge into answers on any topic [19, as recounted in pp. 5–6]. Performances under such manipulation are plainly outside the appropriate range. The same holds for less exotic interference: deliberately misleading or adversarial prompts, including prompt injections, can put a model outside the conditions for which its competences are calibrated. However, compared to human perception, the input channel itself is clean: there is no fog and no bad lighting between prompt and model.

Competences are individuated by mechanisms together with the ranges in which those mechanisms track truth. Not every internal mechanism that sometimes yields truth is a competence with a wide range of appropriateness. This is one place where mechanistic interpretability (MI) becomes helpful: Beckmann and Queloz show that models rely on many different kinds of mechanisms [19, pp. 37–38]. On the one hand, they have sophisticated circuits such as planning the structure of a poem or a greater-than circuit for numerical comparison [19, p. 34]; on the other hand, they rely on cheap circuits such as memorization [19, p. 37]. A working "greater-than circuit" has a wide appropriate range – all numeric comparisons; a memorized answer has only one – that exact case. Both can be competences, but with very differently sized ranges.

To conclude, we should expect many applications to be appropriate for modern LLMs. As demonstrated, however, broad appropriateness cannot just be assumed but must be assessed case by case, which is what we will tackle in the next section.

4.4 C4: Success through Competence

There is an easy case one could be tempted to make. Test two models under identical conditions, one trained, the other not; the first succeeds and the second does not; since no other variable changed, the success must be caused by whatever training installed. And seeing what modern LLMs can do, chance is not a serious rival hypothesis. There is, to be sure, a statistical process going on in the background, but concluding that the whole thing is therefore "not competence but statistics" is a wild move: if we refuse to distinguish the macro level from its micro implementation, we must equally ask whether humans are qualified to possess knowledge at all. However, this argument establishes C1, not C4. That the professional archer outperforms the amateur does not show that each of his hits manifests his competence: sometimes the wind carries a stray arrow in.

Worse, competences can be strange. Suppose an archer's training left him aiming systematically left, while a constant crosswind pushes every arrow right. He hits reliably

– is that competence? Sosa's apparatus answers yes, *relative to those conditions*: the skew-plus-wind pair defines a range in which his disposition makes success highly likely, and removing either element makes him miss. His "knowledge of how to hit" does not transfer to calm conditions with a straight aim, but within its odd range it is genuine – which means C4 hands part of its question back to C3: when a system succeeds via a cheap mechanism, the question is whether the case lay inside that mechanism's appropriately normal range.

That the mechanism is opaque, even to the agent itself, is no obstacle. The chicken sexer reliably sorts chicks without being able to say how, and Sosa counts the resulting beliefs as apt and hence as animal knowledge [14, p. 76]. Knowing how one does it, or even *that* one does it, is not required at the animal level. In the chicken sexer, the mechanism sits *within* the agent; there is a competence, however unconscious. In the same way, whatever mechanisms LLMs employ can be genuine competences relative to their conditions.

Guessing and benchmark-specific shortcuts can be partly excluded behaviorally. Stable success across held-out paraphrases makes lucky guessing implausible, but repetition alone cannot distinguish a narrow deterministic proxy from a broad competence. This is the practical face of the skewed archer, often called "benchmaxxing": a model may perform impressively on a public benchmark because deployment lies outside the proxy mechanism's range. Within the benchmark, such answers can still be apt, but the knowledge does not transfer beyond the reference class for which the mechanism works – which disposition, with which range, made a given success apt is precisely the generality problem familiar from any externalist epistemology.

At this point one may object that all of this is special pleading. With humans, behavioral evidence plus ordinary controls suffices for competence attribution; why demand more of LLMs? The background situation differs: humans have shared etiology, developmental history, and a folk-psychological attribution practice calibrated over millennia, whereas LLMs do not. Principled doubts from the Chinese room [20] onward also leave behavioral evidence underdetermined in a way it is not for humans. Still, C4 is a fact about etiology – a matter of being, not of our knowing it – whether or not anyone inspects the mechanism. MI is therefore not *metaphysically* required for LLM aptness, but it is *epistemically* useful because it reads the etiology directly [19]. With humans, we settle for less because we must.

4.5 Verdict

Some argue that LLMs have at best "unjustified true beliefs," because they possess no mechanisms to validate truth claims [21] or to evaluate whether a belief is warranted

[22]. On an externalist account such as Sosa's, however, no such internal validation is required for animal knowledge. If a system has a competence that makes the truth of its beliefs highly likely in appropriate conditions, and a given belief's truth manifests that competence, then the system knows – even if it does not know that it knows. Since there plausibly exist situations in which LLM beliefs are apt in exactly this sense, LLMs have at least some animal knowledge, conditional on the belief hypothesis of Section 3. The claim is existential and token-level: it awards no blanket status to LLMs, and it is constrained by C3 and C4.

5 Reflective Knowledge in LLMs

LLMs can have some kind of knowledge, but do they also know *that* they know? On the surface, the answer appears to be yes. Ask a modern frontier model whether it knows that $1 + 1 = 2$ or whether Germany is a real country, and it will not only affirm this but express appropriate confidence, cite sources, and give reasons. At first glance, that looks like reflective knowledge. But giving reasons is neither necessary nor sufficient for K^+: reflective knowledge is apt meta-belief, not account-giving. And the reasons themselves could be trained first-order behavior.

Compare two ways a model might produce the sentence "I cannot answer accurately about private events." The first is a reflex: *user mentions a private event → emit the disclaimer*, a pattern installed wholesale in post-training. The second involves an assessment: the model represents the question, finds no relevant information available to it, and reports that finding. The outward token sequence is identical. Output alone therefore underdetermines whether there is a higher-order performance at all, let alone an apt one. Two questions follow: what evidence could distinguish the two mechanisms, and what standard of evidence is it fair to demand?

5.1 The Evidential Standard

Consider a reliable bird recognizer who correctly identifies a bird through well-trained perception but systematically believes that she is only guessing. Her first-order belief may be apt while her higher-order belief is not. Now consider a recognizer who discriminates reliably between favorable and degraded viewing conditions, expresses confidence accordingly, and seeks a closer look when conditions are poor. This pattern is evidence not merely of first-order accuracy but of a competence for assessing it. So human metacognition is often assessed behaviorally: trial-by-trial confidence discrimination, feeling-of-knowing judgments, and information seeking. Comparable paradigms can be used for infants and nonhuman animals who cannot provide verbal reports of their own minds. These experiments must exclude simple cue-based alternatives, but the alternatives are not excluded merely because human-like neural machinery is missing.

Human introspective reports are themselves fallible and often based on plausible causal theories rather than direct access to cognitive processes [23]. If direct neural read-off were required for reflective knowledge, its presence in humans would also be doubtful.

This motivates a principle of evidential parity:

> When a pattern of reliable, counterfactually discriminating, and behaviorally integrated performance is sufficient evidence for a competence in humans or nonhuman animals, a relevantly similar pattern is prima facie evidence for that competence in an artificial system, unless an independently supported difference defeats the inference.

The principle does not say that identical outputs prove identical minds. Human attributions also draw on shared biology, development, and first-person analogy; because LLMs are trained on human linguistic behavior, imitation is a plausible alternative for isolated utterances. Priors may therefore differ. But a different prior does not justify requiring MI only for one substrate. As Cappelen and Dever emphasize, describing a system at a low computational level does not by itself defeat higher-level explanations [4]. The question is whether the alleged alternative explains the whole controlled pattern at least as well. A rule such as *private event* → *disclaimer* can mimic one expression of uncertainty; it becomes implausible when the system discriminates likely success trial by trial, adjusts when evidence enters, and uses the assessment to answer, defer, or seek information.

5.2 The Behavioral Evidence

The best-studied behavioral evidence has exactly this structure. Kadavath et al. asked models both first-order questions and the second-order question whether they know the answers. Larger models turn out to be well calibrated on the probability that their own answers are true, and a trained estimate P(IK) – the probability that "I know" – discriminates answerable from unanswerable questions, partially generalizes to new task types, and rises appropriately when relevant source material or hints toward a solution appear in the context [18]. The meta-signal is sensitive to changes in the *basis* available for the first-order answer, which is what one expects of a meta-competence rather than of a crude disclaimer. Calibration degrades on unfamiliar tasks, which shows that the meta-competence's appropriate range is not unbounded. Arguably, humans are miscalibrated outside their element too. More recent work reports converging indicators: internal belief representations causally drive action selection, and models can monitor and report them [24, p. 1].

Keeling and Street grant that credence attributions to LLMs may well be literally true, but argue that the elicitation techniques used to measure them may fail to track

that truth [25]. Singh, Linzen, and Ravfogel re-examine two introspection paradigms – whether models can distinguish tampering with their internal states from mere input manipulations, and whether models can predict labels derived from their own hidden states – and find both underdetermined: input-only classifiers match the models' self-predictions, and performance falls toward chance under relabeled controls [26]. These results undermine strong, privileged introspection, not all meta-competence: human introspective reports are fallible, and Sosa requires no privileged read-off. An apt meta-belief may be produced through reliable indirect cues; the negative results narrow which mechanisms can be doing the work and defeat overclaims.

A further objection targets Sosa's framework more specifically: P("I know") and confidence track truth or answerability, whereas the content of the reflective belief must be *aptness* (R3) – not identical. Two mitigations: the ways a first-order LLM belief fails to be apt (no basis in the training distribution, contamination, ...) are largely what these signals are sensitive to, so a state discriminating "I have a competent basis here" from "I do not" is close in content to a belief about aptness even if the word never occurs; and human higher-order judgment likewise runs on fallible proxies such as fluency and familiarity [23] – if a proxy reliably tracks competent success within its proper range, its outputs plausibly qualify as apt meta-beliefs, and here too Sosa's externalism cuts in the model's favor.

5.3 The Etiological Evidence

Behavior is one kind of evidence for a meta-competence; the training process that would have to produce it is another. Meta-competence should resemble first-order competence in this respect: we should expect to find only what was trained for. In post-training, models are penalized for falsehoods. Systematically avoiding falsehoods they would otherwise produce requires something in the model to discriminate likely-true from likely-false outputs, so the training pressure selects for the disposition R2 demands. This is not first-order accuracy smuggled into the second order: calibrated hedging and refusal are actions guided by an assessment of first-order prospects, not better first-order answers. P(IK) is precisely a trained meta-signal [18], and comparing post-trained models with their bases, Wes, Nicholas, and Jack find traces of the Assistant monitoring its own behavior in the model's internal workspace [27].

Where should this meta-competence stop? Where the training signal stops tracking truth. Bad training data is one boundary: a model fed misinformation will confidently repeat it, as a misinformed human would. Human preference feedback is a subtler one: raters sometimes prefer confident, agreeable, or flattering answers over true ones, at least in domains they cannot check [28]. The meta-competence should therefore be gradient: strongest where the reward was verifiable (math, code, tool use), weaker where it tracked

human approval – C3 again, one level up, set by the verifiability of the training signal. Section 4's defeaters (internal tampering, adversarial prompts, ...) apply here too.

5.4 What Mechanistic Evidence Adds

If MI is not required to assess the question, what is it for? It can *corroborate* a behavioral attribution by finding a suitable mechanism; *localize* the competence by identifying its domain and failure conditions; and *defeat* the attribution where the behavior turns out to be generated by a route that does not track what it seems to track.

Recent work makes the positive architecture at least plausible. Wes, Nicholas, and Jack report that verbalizable representations form a global-workspace-like band in language models: internal contents that are reportable, flexibly available, and causally usable across tasks [27], paralleling the global neuronal workspace, with important disanalogies [29]. The evidence is moderate: injected concepts are reported in the majority of trials, and swapping workspace vectors redirects multi-hop answers in 54–70% of trials, while activation injections may represent abnormal C3 conditions [27]. In a non-peer-reviewed research note, agam_bhatia, camilablank, and Nanda find a "hedging" representation that fires before the model qualifies an open-ended answer, and whose suppression makes it commit to a single option [30]. This is the kind of substrate a meta-competence needs, though it is not yet a demonstrated belief with the content "my first-order belief is apt." Mechanistic evidence can also cut the other way: Beckmann and Queloz's finding that models exploit a motley mix of heterogeneous mechanisms in parallel [19] means that any meta-competence may itself be a patchwork, apt in some regions of its domain and absent in others. MI then bounds the competence more than proving it in general.

There is also a principled reason not to promote MI to a gatekeeping role: any mechanistic criterion we might choose – a dedicated meta-circuit, later-layer read-off, causal control by the meta-representation – is a stipulation Sosa's account does not contain, and one humans have never been shown to satisfy either; under the same microscope, human meta-judgment might itself resolve into indirect cues and post hoc theorizing [23], which would prove too much by stripping humans of reflective knowledge too. Evidential parity runs in both directions: the standards must be demanding, and the same for both.

5.5 Verdict

Behaviorally, models discriminate what they know, with sensitivity to the available basis [18, 24]. Etiologically, post-training selects for exactly such a discriminating disposition and even leaves visible self-monitoring traces [27]. Mechanistically, causally efficacious representations of the model's own processing exist, from the workspace band to hedging tokens [27, 30]. The negative results – confounded introspection tests [26], a motley mix of underlying mechanisms [19] – have recognizable human analogues [23][19, p. 37] and

target requirements that Sosa's externalism does not impose.

Where does this leave $K^+p \leftrightarrow KKp$? Conditional on functional belief, the combined evidence supports a defeasible attribution of reflective knowledge in some contexts: where a model's first-order successes are independently creditable to a competence, and its knowledge-judgments reliably discriminate those successes from failures with sensitivity to the available basis, the best cases satisfy R1–R4. The attribution could be overturned by more evidence, such as contamination, a first-order imitation explanation, or results in the style of Singh, Linzen, and Ravfogel. However, defeasibility is the normal condition of competence attribution everywhere, not a reason to withhold it until the mechanism is perfectly understood.

6 Conclusion

This paper asked whether LLMs can know, and know that they know, in Sosa's sense. The answer is conditional throughout on the working hypothesis of Section 3: that some LLM states qualify as beliefs, at least in Sosa's minimal sense. Given that hypothesis, two results.

First, at least some LLM performances plausibly constitute animal knowledge. That result is not trivial. It is carried by externalism – under a justified-true-belief conception it might not follow – and it is disciplined by C3 and C4. The generality problem does not disappear, but LLMs are the first knowers for whom it is empirically tractable: mechanistic interpretability can, in principle, read the etiology of a performance directly, something we have never been able to do for one another.

Second, on the reflective question I have defended an evidential rather than an architectural standard. The same behavioral criteria used to attribute metacognition to humans and animals (discrimination, counterfactual sensitivity to the available basis, integration with action) support a defeasible attribution of reflective knowledge to current LLMs. Mechanistic evidence corroborates, localizes, or defeats such attributions; it is not their precondition, and any mechanistic criteria strict enough to exclude LLMs by stipulation would likely exclude humans too.

Both verdicts should be understood as token-level and graded. LLMs do not know everything they say – but neither do we, and Sosa's framework was built precisely to assess performances one at a time rather than to award or deny an epistemic status to a whole kind of mind. This suggests a practical research program: using Sosa's framework and MI to map, competence by competence, where a model's successes manifest a reliable competence and where it does not. Knowing where a model's competences end is knowing where our trust should end too.

References

[1] Plato. *Plato: Complete Works*. Hackett Publishing, May 1, 1997. 1840 pp. ISBN: 978-1-60384-670-7. Google Books: fMhgDwAAQBAJ.

[2] Sha Sajadieh et al. *Artificial Intelligence Index Report 2026*. June 29, 2026. DOI: 10.48550/arXiv.2606.15708. arXiv: 2606.15708 [cs.AI]. URL: http://arxiv.org/abs/2606.15708 (visited on 07/15/2026). Pre-published.

[3] Raphaël Millière and Cameron Buckner. "The Philosophy of Language Models". In: *Philosophy Compass* 21.3 (2026), e70095. ISSN: 1747-9991. DOI: 10.1111/phc3.70095. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/phc3.70095 (visited on 07/08/2026).

[4] Herman Cappelen and Josh Dever. *Going Whole Hog - A Philosophical Defense of AI Cognition*. URL: https://philpapers.org/archive/CAPGWH.pdf#page=100.55 (visited on 05/09/2026). Pre-published.

[5] Matthias Steup and Ram Neta. "Epistemology". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2025. Metaphysics Research Lab, Stanford University, 2025. URL: https://plato.stanford.edu/archives/fall2025/entries/epistemology/ (visited on 07/12/2026).

[6] Ernest Sosa and Ernest Sosa. *A Virtue Epistemology: Apt Belief and Reflective Knowledge, Volume I*. Oxford, New York: Oxford University Press, June 28, 2007. 164 pp. ISBN: 978-0-19-929702-3.

[7] Emma Borg. "LLMs, Turing Tests and Chinese Rooms: The Prospects for Meaning in Large Language Models". In: *Inquiry* (Jan. 7, 2025), pp. 1–31. ISSN: 0020-174X, 1502-3923. DOI: 10.1080/0020174X.2024.2446241. URL: https://www.tandfonline.com/doi/full/10.1080/0020174X.2024.2446241 (visited on 04/18/2026).

[8] Daniel A. Herrmann and Benjamin A. Levinstein. "Standards for Belief Representations in LLMs". In: *Minds and Machines* 35.1 (Dec. 7, 2024), p. 5. ISSN: 1572-8641. DOI: 10.1007/s11023-024-09709-6. arXiv: 2405.21030 [cs.AI]. URL: http://arxiv.org/abs/2405.21030 (visited on 07/12/2026).

[9] B. A. Levinstein and Daniel A. Herrmann. "Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks". In: *Philosophical Studies* 182.7 (July 2025), pp. 1539–1565. ISSN: 0031-8116, 1573-0883. DOI: 10.1007/s11098-023-02094-3. arXiv: 2307.00175 [cs.CL]. URL: http://arxiv.org/abs/2307.00175 (visited on 07/08/2026).

[10] Emily M. Bender et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? " In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, Mar. 1, 2021, pp. 610–623. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445922. URL: https://dl.acm.org/doi/10.1145/3442188.3445922 (visited on 05/06/2026).

[11] Murray Shanahan. "Talking about Large Language Models". In: *Communications of the ACM* 67.2 (Jan. 25, 2024), pp. 68–79. ISSN: 0001-0782. DOI: 10.1145/3624724. URL: https://dl.acm.org/doi/10.1145/3624724 (visited on 07/15/2026).

[12] Simon Goldstein and Benjamin A. Levinstein. *Does ChatGPT Have a Mind?* June 27, 2024. DOI: 10.48550/arXiv.2407.11015. arXiv: 2407.11015 [cs.CL]. URL: http://arxiv.org/abs/2407.11015 (visited on 07/15/2026). Pre-published.

[13] David J. Chalmers. *Propositional Interpretability in Artificial Intelligence*. Jan. 27, 2025. DOI: 10.48550/arXiv.2501.15740. arXiv: 2501.15740 [cs.AI]. URL: http://arxiv.org/abs/2501.15740 (visited on 07/15/2026). Pre-published.

[14] Ernest Sosa. *Judgment and Agency*. Oxford University Press, Apr. 1, 2015. ISBN: 978-0-19-871969-4. DOI: 10.1093/acprof:oso/9780198719694.001.0001. URL: https://doi.org/10.1093/acprof:oso/9780198719694.001.0001 (visited on 07/12/2026).

[15] Wayne Xin Zhao et al. "A Survey of Large Language Models". In: *Frontiers of Computer Science* 20.12 (Dec. 2026), p. 2012627. ISSN: 2095-2228, 2095-2236. DOI: 10.1007/s11704-026-60308-3. arXiv: 2303.18223 [cs.CL]. URL: http://arxiv.org/abs/2303.18223 (visited on 08/04/2026).

[16] Long Ouyang et al. *Training Language Models to Follow Instructions with Human Feedback*. Mar. 4, 2022. DOI: 10.48550/arXiv.2203.02155. arXiv: 2203.02155 [cs.CL]. URL: http://arxiv.org/abs/2203.02155 (visited on 08/04/2026). Pre-published.

[17] SimpleBench Team. *SimpleBench*. URL: https://simple-bench.com/simple-bench.com (visited on 07/12/2026).

[18] Saurav Kadavath et al. *Language Models (Mostly) Know What They Know*. Nov. 21, 2022. DOI: 10.48550/arXiv.2207.05221. arXiv: 2207.05221 [cs.CL]. URL: http://arxiv.org/abs/2207.05221 (visited on 08/04/2026). Pre-published.

[19] Pierre Beckmann and Matthieu Queloz. *Mechanistic Indicators of Understanding in Large Language Models*. Feb. 25, 2026. DOI: 10.48550/arXiv.2507.08017. arXiv: 2507.08017 [cs]. URL: http://arxiv.org/abs/2507.08017 (visited on 05/09/2026). Pre-published.

[20] John R Searle. "Minds, Brains, and Programs". In: ().

[21] Rafael Alvarado. *What Large Language Models Know*. School of Data Science. URL: https://datascience.virginia.edu/news/what-large-language-models-know (visited on 07/12/2026).

[22] Earl Woodruff and Jim Hewitt. "Epistemic Agency in the Age of Large Language Models: Design Principles for Knowledge-Building AI". In: *AI* 7.3 (Mar. 2026), p. 99. ISSN: 2673-2688. DOI: 10.3390/ai7030099. URL: https://www.mdpi.com/2673-2688/7/3/99 (visited on 07/12/2026).

[23] Richard E. Nisbett and Timothy D. Wilson. "Telling More than We Can Know: Verbal Reports on Mental Processes". In: *Psychological Review* 84.3 (1977), pp. 231–259. ISSN: 1939-1471. DOI: 10.1037/0033-295X.84.3.231.

[24] Noam Steinmetz Yalon et al. *Indications of Belief-Guided Agency and Meta-Cognitive Monitoring in Large Language Models*. Feb. 2, 2026. DOI: 10.48550/arXiv.2602.02467. arXiv: 2602.02467 [cs.CL]. URL: http://arxiv.org/abs/2602.02467 (visited on 07/20/2026). Pre-published.

[25] Geoff Keeling and Winnie Street. "On the Attribution of Confidence to Large Language Models". In: *Inquiry* 69.6 (July 3, 2026), pp. 2918–2944. ISSN: 0020-174X, 1502-3923. DOI: 10.1080/0020174X.2025.2450598. arXiv: 2407.08388 [cs.AI]. URL: http://arxiv.org/abs/2407.08388 (visited on 07/20/2026).

[26] Shashwat Singh, Tal Linzen, and Shauli Ravfogel. *Can LLMs Introspect? A Reality Check*. May 25, 2026. DOI: 10.48550/arXiv.2605.26242. arXiv: 2605.26242 [cs.AI]. URL: http://arxiv.org/abs/2605.26242 (visited on 07/20/2026). Pre-published.

[27] Gurnee Wes, Sofroniew Nicholas, and Lindsey Jack. *Verbalizable Representations Form a Global Workspace in Language Models*. URL: https://transformer-circuits.pub/2026/workspace/index.html (visited on 07/08/2026).

[28] Mrinank Sharma et al. *Towards Understanding Sycophancy in Language Models*. May 10, 2025. DOI: 10.48550/arXiv.2310.13548. arXiv: 2310.13548 [cs.CL]. URL: http://arxiv.org/abs/2310.13548 (visited on 08/04/2026). Pre-published.

[29] Stanislas Dehaene et al. "External Commentary on Verbalizable Representations Form a Global Workspace in Language Models". In: ().

[30] agam_bhatia, camilablank, and Neel Nanda. "Towards Surfacing Model Algorithms with Meta-Tokens in the J-Space — AI Alignment Forum". In: (July 20, 2026). URL: https://www.alignmentforum.org/posts/6ek6n7yZ5DzfarJHy/towards-

surfacing-model-algorithms-with-meta-tokens-in-the-j (visited on 07/22/2026).