Artificial Intelligence as an Epistemic Authority? Applying Goldman's Criteria to LLMs

Jonas Strabel

April 10, 2026

1 Introduction

Today's world contains more knowledge than any single person can possess, so specialization is vital. However, this means that one must trust others when it comes to knowledge in domains outside one's own expertise. From an epistemic point of view, this raises mainly two types of questions: (1) how do I know whom to trust, and (2) what should I do with their judgment (e.g. preempt or merely strongly consider it)? While both are certainly interesting, this article will focus on the first question. There is already a large body of literature on this issue, but recently, with contemporary large language models (LLMs), such as Claude 4.6 (Anthropic 2026a) and ChatGPT 5.2 (OpenAI 2026b), becoming much more powerful, a new question has to be answered: can an LLM also be something we can trust in an epistemically relevant sense?

This paper first argues that LLMs can, in principle, qualify as epistemic authorities and then asks under what conditions a non-expert is justified in treating a particular LLM as one.

2 Epistemic Authorities

Epistemic authorities (EAs) are not defined uniformly in the literature, but a traditional idea is that EA is an epistemic authority for S just in case EA is epistemically superior to S, that is, has substantially more true beliefs and fewer false beliefs than S in some domain D. Applied to AI, however, this formulation invites the further question of whether AI systems literally have beliefs. Whether LLMs have beliefs or knowledge in any philosophically robust sense is a genuine open question, but one that this paper deliberately sets aside. For present purposes, it is therefore useful to adopt Hauswald's indicator-based account, which avoids commitment to the language of belief. In paraphrased form, Hauswald's proposal is this:

> EA is a propositional epistemic authority for S with respect to domain D iff, relative to S, the facts used as indicators of truth about EA correctly track the truth of the relevant propositions in sufficiently many cases, and do not do so incorrectly in a disproportionately large number of cases (Hauswald 2024, p. 69).

Here, the outputs of an LLM shall be considered truth indicators. LLMs are trained on human-produced materials and thus inherit, filter, compress, and recombine large bodies of human testimony. Whether this makes them agents in their own right can remain open. For present purposes, what matters is that their outputs can function as truth indicators.

3 Refining the Question and Methodology

Since LLMs output text that can function as a truth indicator, we shall say that LLMs can qualify as EAs. Deference to such an EA then means

> "taking appropriate facts about that EA as strong indicators of the truth of certain propositions" (Hauswald 2025, p. 722).

In the case of LLMs, an appropriate fact would be its output to an appropriate prompt. Hauswald argues for this wider definition not only because it accommodates artificial EAs, but also because it is already necessary in order to accommodate subjects who defer to human EAs in non-testimonial ways or to entities without beliefs, such as scientific communities (Hauswald 2025, p. 722).

The notion of EA is relative: EA may be an epistemic authority for S without being one for another subject S_2. Similarly, EA may be an epistemic authority for S in a domain D, but not in a domain D_2. Whether an LLM really is an EA therefore depends on how well it correctly tracks the truth of relevant propositions in D relative to S. LLMs are thus not EAs simpliciter, but rather can be EAs for a subject S in a domain D.

Just as with human EAs, it is therefore necessary to examine whether a certain LLM is an EA for a certain person in a certain domain. Doing so is not straightforward, since the one who is epistemically inferior cannot directly assess who is superior, because this would require sufficient knowledge in the domain. This issue has been discussed in the literature as the novice-expert problem. To address it, Goldman has proposed his famous five sources of evidence that S might have for trusting a putative expert. While these sources are used to identify human experts, they should also be helpful for assessing the closely related notion of epistemic authority.

In the following section, Goldman's five sources (Goldman 2001, p. 93) will be discussed by looking at (a) whether, and to what extent, they are applicable to LLMs and (b) how they

apply to an ordinary non-expert S in an established domain such as medicine or mathematics D, in order to assess whether trusting the LLM as an EA is rational in such a case. For anyone seeking to assess whether an LLM is an EA for them in a certain domain, the following methodology may serve as guidance. This paper focuses on the ordinary non-expert. A separate question — how experts themselves should treat LLMs as potential authorities at the edge of current knowledge, where the risk of hallucination is highest — is worth examining, but lies outside the present scope.

4 Goldman's Five Sources to Assess Trustworthiness

4.1 A: Arguments Presented by the Contending Experts to Support Their Own Views and Critique Their Rivals' Views

Goldman concedes that this type of evidence may or may not be available, as there might be no publicly accessible defenses or arguments from an expert, or only truncated ones. Even when such material is available, it may not be possible for the novice (the one assessing whether someone is an expert or, in this case, an EA) to evaluate or fully comprehend the evidence. However, sometimes it is available, and there are indirect factors for assessing which expert is superior.

This paper argues that the situation is similar with LLMs. In some domains, there are sources on the internet in which a supposed expert argues with or tests an LLM in a way that could be considered to fit this category of Goldman (e.g. *Bernie vs. Claude* 2026). One area in which the LLM has an advantage is that the novice can directly question it and ask for simpler explanations that they can understand, in order to assess whether what the LLM says is at least coherent and plausible. Obviously, because the person is a novice, they may not be able to verify what is being said, but logical contradictions, incorrect citations, or other inconsistencies can sometimes be identified, and these often reveal the model's weakness.

In some domains, looking at non-testimonial indicators might also be helpful. In coding, for example, top-tier companies say that most of the code in their company is written by LLMs Nolan 2026. If software engineers, who are experts in their field, use code produced by an LLM instead of writing it entirely themselves, this indicates that the LLM is at least somewhat competent in that field. Certainly, other factors such as time savings also play a role, but this would not be possible if the LLM lacked any competence in that area. Similarly, many doctors rely on other forms of AI in their judgments.

Another consideration relevant to this paper is whether Goldman's category might be understood somewhat differently here. Since the question is not who the best expert is, but rather *Is this LLM an EA for me?*, direct debate between the novice and the LLM might itself be

an indicator. Many users of powerful LLMs have likely noticed that LLMs are dialectically strong in many domains. When presented with an argument for a position, the LLM often appears to have encountered such an argument already and to have counterarguments readily available. Goldman considers this an indicator of superiority in the domain Goldman 2001, p. 95.

A caveat is worth noting here. For human experts, dialectical strength and truth-tracking tend to be coupled: a well-argued position usually reflects genuine domain competence. For LLMs, however, fluency and accuracy are partially decoupled — a model can produce a highly coherent, well-structured rebuttal that is nonetheless factually wrong. The novice should therefore treat dialectical performance as a weak indicator at best: useful for ruling out obvious incompetence, but insufficient on its own.

4.2 B: Agreement from Other Experts: The Question of Numbers

In this section, Goldman shows why a larger number of experts agreeing with an expert E is only meaningful if they are conditionally independent, i.e. if they came to believe the conclusion independently and do not simply do so because E believes it. A greater number of agreements is a good indicator only if there is a genuine consensus.

Applying this to LLMs is not completely straightforward. The issue is best discussed by distinguishing two cases.

Established Knowledge For established knowledge, LLMs are not additional experts who push the numbers in one direction or the other, but are in a sense a proxy for human experts. This is because an LLM does not output something because it made its own observations, but because of what other humans have claimed in the data on which it was trained. It is therefore certainly not independent. But that does not mean that this point cannot help to establish LLMs as EAs. Where consensus exists, LLMs often reflect that consensus in their output. Where no consensus exists, an LLM can often present the competing positions well and will usually respond with nuance.

There are two things to keep in mind. One is question framing. When a question is framed in a way that favors one conclusion over another or reveals certain affiliations, answers may differ from what a neutral question would have produced. For example, a question framed as 'is it plausible that X?' tends to elicit a more permissive answer than 'which is more likely, X or not-X?' — even when the underlying evidence is the same. Someone prompting 'is it plausible that X?' might get the impression that the LLM supports X, whereas if they had asked 'which is more likely, X or not-X?', the LLM might have answered 'not-X'. It is worth noting, however, that when it comes to questions for which a clear consensus exists, such as climate change, current LLMs generally do not treat the contrary view as equally plausible, which is reassuring.

The second thing is bias in training. Since the model is trained primarily in the West on largely Western data and within largely Western normative frameworks, it reflects those influences. This does not by itself show that its outputs are false, but it is worth noting that the model reflects the perspective of the context in which it was trained. It is also possible that, in the future, malicious actors could influence models in ways that distort scientific facts or even deny established consensus.

However, as of now, it seems that the models are trained on a sufficiently broad range of data to reflect consensus where one exists and to explain multiple positions where none does. This means that, for many domains in which sufficient data are available, LLMs may perform at something like an expert level with respect to established knowledge, precisely because they proxy the judgments of human experts. This speaks in favor of accepting the LLM as an EA, provided one is not oneself an expert in that domain.

New and Transferred Knowledge A more difficult case is that in which LLMs are used to generate new knowledge or to apply knowledge and skills in novel ways. As models become more powerful, this has started to occur. For example, GPT 5.2 derived a new result in theoretical physics (OpenAI 2026c), doctors use AI to evaluate images, and software engineers use it to write novel source code.

According to the definition of EA above, established knowledge would suffice to make an LLM an EA in a domain D. But does this mean that one should also trust it with respect to new knowledge that is not merely a proxy for what experts have written? In this case, judgments from human experts would certainly help to establish credibility, because we do not know how an LLM arrives at its conclusions — certainly not in the same way as a human expert, and not simply because a human said it. That is, of course, assuming that the output is not manipulated through a biased prompt. In practice, new knowledge produced by an LLM is still typically evaluated by human experts when it really matters, such as when a doctor reviews an image analysis or a software engineer reviews generated code. There is growing evidence that in some fields (such as code generation), LLMs are becoming strong enough that parts of their novel output are often reliable. Whether LLMs can be considered EAs with respect to genuinely new knowledge will remain an open question in this paper, because of the complexity of the issue.

4.3 C: Appraisals by "Meta-Experts" of the Experts' Expertise (Including Appraisals Reflected in Formal Credentials Earned by the Experts)

Another appeal to expert judgment is Goldman's category C. Here, meta-experts, certifications, ratings, credentials, and other such indicators are considered. He argues that if meta-experts rate someone highly, should this not be a reason to rely on them?

He does mention that he treats ratings and credentials as signaling 'agreement' with other experts, because he assumes that established authorities certify trainees as competent when they demonstrate (1) mastery of the methods fundamental to the field and (2) knowledge of, or belief in, propositions that are deemed to be fundamental facts or laws of the discipline. Whether this applies, or can apply, to an LLM is a difficult question and again lies outside the scope of this paper. This places a limitation on trusting the LLM as an EA, since this is an important basis for being trusted with making new discoveries in a field.

Despite LLMs not having degrees or credentials in the traditional sense, this category can be applied to them in a different way: benchmarks and tests. Academia and industry have developed many different benchmarks and tests to evaluate model performance, compare models, and measure progress.

The creators of the MMLU benchmark estimated human expert performance at 89.8% Hendrycks et al. 2021, while ChatGPT-4o, a model that has since been surpassed, scored 87.4% (OpenAI 2026d). Note that the 89.8% figure is an estimate for experts in their respective fields. This test was criticized, however, for not accurately predicting real-world performance, and a new test was subsequently introduced (Wang et al. 2024). GPQA is another benchmark created to test LLMs. Holders or pursuers of PhDs score 65% or 74% (after reconsideration) on the test, while frontier models like Opus 4.6 and Gemini 3 Pro reach 91% and 93%, respectively. This shows that LLM performance has matched or even surpassed that of human experts in certain respects. While these models have not surpassed the very top experts in a field, they certainly do so relative to the average human (non-experts reached only 34% accuracy on GPQA, far below current LLMs). These tests span many domains and scientific disciplines, from history, philosophy, and law to physics and mathematics. Beginning with GPT-4 in 2023, LLMs have also passed law and medicine entry exams (Katz et al. 2026; Nori 2026). When one is concerned not with problem-solving but with up-to-date facts, however, one has to be somewhat more careful. Google's FACTS benchmark evaluates LLM performance on facts about people, celebrities, politics, and related topics (Google 2025). The best model (Google's Gemini 3 Pro) achieves 68.8%.

Apart from benchmarks, experts in various domains often comment on model performance. To take another example from software engineering, some experienced engineers concede that the model writes code better than they themselves do, and that they rarely write code anymore.

Clearly, in many domains, models perform exceptionally well — better than most humans. In these domains and for these tasks, this speaks in favor of accepting LLMs as EAs for those who are not experts in the relevant field. In some tasks, such as recent facts about celebrities or politics, there is still room for improvement. To find out whether one can trust an LLM in a certain domain or for a certain task, one can consult benchmarks or tests conducted for it or for adjacent domains. Just as one would check whether a person has a specific degree or relevant experience in a field, one can check whether a model has performed well on a relevant benchmark, passed an entrance exam, or been appraised positively by other experts. If it has, one may be justified in accepting the LLM as an EA in that domain. It

is, however, worth noting that benchmarks do not always reflect real-world performance. It is therefore best not to rely on a single benchmark, but rather on a combination of several, together with appraisals from practicing experts.

4.4 D: Evidence of the Experts' Interests and Biases Vis-à-Vis the Question at Issue

Goldman argues that the presence of bias justifies preferring the expert without bias over the one with a bias. Lying is a common form of bias, but it can also be subtler, making an expert's opinion less likely to be accurate even if they are sincere. Group bias or whole-discipline bias may be even more difficult to detect.

Since the LLM itself does not have self-interest in the human sense (at least not yet, as far as we know), it is not subject to individual human bias in the way a person may be through, for example, industry funding. A complication here is that LLMs are designed and deployed by companies that do have interests — commercial, reputational, and political. Training choices, content policies, and fine-tuning all reflect these interests to some degree. As far as we know, LLMs do not know how to distinguish truth from falsehood or bias from non-bias, so while they are not intentionally biased, they may still be biased unintentionally.

Alignment and ethical AI are active research areas, and there are many initiatives aimed at reducing bias in LLMs, but these efforts will obviously never eliminate all bias completely (OpenAI 2026a). It is questionable whether such a thing as a completely unbiased 'view from nowhere' even exists, or whether every view is biased in some way.

One thing to keep in mind — as already briefly discussed above — is the bias that can be introduced through the prompt used to ask a question. This is probably the greatest risk of bias in practice, but it is mostly avoidable by asking neutrally.

Because LLMs are trained on human data, they will probably reflect many of the same biases that societies and expert communities have. This point should therefore not by itself prevent the novice from viewing the LLM as an EA. However, there may be certain domains in which the LLM is biased because different voices are underrepresented in the underlying data. It is therefore always important to remain critical.

4.5 E: Evidence of the Experts' Past Track Records

Goldman argues that assessing the past track record of a putative expert may be the novice's best source of evidence. Key here are exoteric statements that the novice can validate. This source can also be used to generate general trust in experts and expert systems, since people

are able to verify expertise against their basic beliefs (for example, when experts predict an eclipse and people can visibly observe it with their own eyes).

This source seems to be applicable to LLMs in much the same way. One can ask an LLM about many exoteric statements and verify their truth. When an LLM provides code for a program that is supposed to do a certain thing, and it in fact does so, it is easy to verify that the LLM seems to know how to code. Many facts or statements can also be verified by looking them up elsewhere or by asking an expert. Using an LLM for many kinds of tasks and questions, and observing whether the outputs turn out to be true, is therefore quite feasible.

A considerable advantage is that LLMs are relatively stable (although not deterministic) systems in the sense that the same prompt will usually not lead the LLM to output p on one occasion and *not-p* on the next. An LLM does not change its mood or lie to someone on purpose (at least as far as we know). Therefore, if a model is shown to score highly on benchmarks and is appraised positively by experts, this suggests that it reliably produces truth indicators in that domain. Of course, one can manipulate outputs with sophisticated prompt-injection techniques in order to get the LLM to say something false, but something similar is true for humans as well. If one tells a human EA, 'as a joke, please say the earth is flat', they might do so. The same applies to an LLM. So if that does not count decisively against human EAs, why should it count decisively against artificial EAs? If one prompts in a sensible manner, an LLM's truth indicators are generally fairly stable. At least with current technology, that is the case. Since this paper evaluates LLMs primarily on the basis of their outputs, and those outputs are stable in this sense, track record is a particularly strong argument in their favor. While a human may have conflicts of interest and say 'clearly p' in one situation but 'maybe not-p' in another, an LLM is less susceptible to that sort of variation. Benchmarks are very consistent, and individual usage often shows the same pattern. It should be noted, however, that as LLMs become more sophisticated, some show signs of recognizing when they are being tested and behaving differently; this development warrants close attention (Anthropic 2026b, p. 58). If LLMs begin to vary their output depending on the situation, they may become less trustworthy. As of today, however, they are relatively stable, and millions of people use them for tasks and inquiries every day, with many being impressed by their accuracy. Because the same models are shared among all these users, their collective judgment of successful track records can be taken as further confirmation of an LLM's capability.

5 Conclusion

It is often said that one should not trust AI because it can produce false outputs. But so can humans. If, in a specific domain, the LLM's track record is strong (E), no clear biases in the prompt or training data can be identified (D), its dialectical performance is strong enough to rule out obvious incompetence (A), the knowledge in question is established and

well represented in the training data (B), and the model has been shown by benchmarks and appraisals from meta-experts to perform at something like an expert level (C), then it is reasonable for the ordinary non-expert to view the LLM as an EA. The model may still say something false, but so may human EAs. The reason we trust EAs over our own judgment is that, on average, we fare better with this strategy. After evaluating LLMs functionally — that is, according to their outputs — it seems that in certain domains, and for non-expert users dealing with established knowledge, this strategy can also apply to artificial EAs.

A question that merits closer examination is how experts themselves should treat artificial EAs. While this framework and its conclusion may in principle also apply to expert-LLM relationships, that question deserves further analysis, since LLMs tend to hallucinate confidently when it comes to things they do not know, and experts usually work at the very edge of the knowledge available to LLMs.

This paper's evaluation rests on the current state of LLMs. If, in the future, they change or improve significantly, or if much more is understood about how they work internally, certain conclusions could change. This can be examined in future work. However, it is not implausible that the central conclusion will remain broadly intact. If it shifts in any direction, it will probably shift toward favoring LLMs as EAs even more, as they continue to improve.

References

Anthropic (2026a). *Claude Opus 4.6*. en. URL: https://www.anthropic.com/news/claude-opus-4-6 (visited on 02/16/2026).

— (2026b). *Claude Sonnet 4.5 System Card*. URL: https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf#page=59&zoom=100,96,182 (visited on 02/16/2026).

Bernie vs. Claude (Mar. 2026). URL: https://www.youtube.com/watch?v=h3AtWdeu_G0 (visited on 04/09/2026).

Goldman, Alvin I. (2001). "Experts: Which Ones Should You Trust?" In: *Philosophy and Phenomenological Research* 63.1, pp. 85–110. ISSN: 0031-8205. DOI: 10.2307/3071090. URL: https://www.jstor.org/stable/3071090 (visited on 04/10/2026).

Google (Dec. 2025). *FACTS Benchmark Suite: a new way to systematically evaluate LLMs factuality*. en. Section: Responsibility & Safety. URL: https://deepmind.google/blog/facts-benchmark-suite-systematically-evaluating-the-factuality-of-large-language-models/ (visited on 02/16/2026).

Hauswald, Rico (2024). *Epistemische Autoritäten: Individuelle und plurale*. de. Berlin, Heidelberg: Springer. ISBN: 978-3-662-68749-9 978-3-662-68750-5. DOI: 10.1007/978-3-662-68750-5. URL: https://link.springer.com/10.1007/978-3-662-68750-5 (visited on 02/16/2026).

— (Nov. 2025). "Artificial Epistemic Authorities". In: *Social Epistemology* 39.6. _eprint: https://doi.org/10.1080/02691728.2025.2449602, pp. 716–725. ISSN: 0269-1728. DOI: 10.

1080/02691728.2025.2449602. URL: https://doi.org/10.1080/02691728.2025.2449602 (visited on 02/16/2026).

Hendrycks, Dan et al. (Jan. 2021). *Measuring Massive Multitask Language Understanding*. arXiv:2009.03300 [cs]. DOI: 10.48550/arXiv.2009.03300. URL: http://arxiv.org/abs/2009.03300 (visited on 02/16/2026).

Katz, Daniel Martin et al. (2026). "GPT-4 passes the bar exam". In: *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 382.2270 (), p. 20230254. ISSN: 1364-503X. DOI: 10.1098/rsta.2023.0254. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10894685/ (visited on 02/16/2026).

Nolan, Beatrice (2026). *Top engineers at Anthropic, OpenAI say AI now writes 100% of their code—with big implications for the future of software development jobs*. en. URL: https://fortune.com/2026/01/29/100-percent-of-code-at-anthropic-and-openai-is-now-ai-written-boris-cherny-roon/ (visited on 04/09/2026).

Nori, Harsha (2026). *Capabilities of GPT-4 on Medical Challen...* URL: https://axi.lims.ac.uk/paper/2303.13375 (visited on 02/16/2026).

OpenAI (Mar. 2026a). *Defining and evaluating political bias in LLMs*. en-US. URL: https://openai.com/index/defining-and-evaluating-political-bias-in-llms/ (visited on 04/09/2026).

— (Apr. 2026b). *Entdecke GPT-5.2*. de-DE. URL: https://openai.com/de-DE/index/introducing-gpt-5-2/ (visited on 04/09/2026).

— (Feb. 2026c). *GPT-5.2 derives a new result in theoretical physics*. en-US. URL: https://openai.com/index/new-result-theoretical-physics/ (visited on 02/16/2026).

— (2026d). *Hallo GPT-4o*. de-DE. URL: https://openai.com/de-DE/index/hello-gpt-4o/ (visited on 04/10/2026).

Wang, Yubo et al. (Nov. 2024). *MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark*. arXiv:2406.01574 [cs]. DOI: 10.48550/arXiv.2406.01574. URL: http://arxiv.org/abs/2406.01574 (visited on 02/16/2026).