aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Current Directions in Psychological Science
1–9

DOI: 10.1177/09637214251391158
www.psychologicalscience.org/CDPS
Sage

Metacognition and Uncertainty Communication in Humans and Large Language Models

Mark Steyvers[1,2] **and Megan A. K. Peters**[1,2,3,4]
[1]Department of Cognitive Sciences, University of California, Irvine;
[2]Center for the Neurobiology of Learning and Memory, University of California, Irvine;
[3]Center for Theoretical Behavioral Sciences, University of California, Irvine; and
[4]Brain, Mind & Consciousness, Canadian Institute for Advanced Research, Toronto, Ontario, Canada

Abstract

Metacognition—the capacity to monitor and evaluate one's own knowledge and performance—is foundational to human decision-making, learning, and communication. As large language models (LLMs) become increasingly embedded in both high-stakes and widespread low-stakes contexts, it is important to assess whether, how, and to what extent they exhibit metacognitive abilities. Here, we provide an overview of the current knowledge of LLMs' metacognitive capacities, how they might be studied, and how they relate to our knowledge of metacognition in humans. We show that although humans and LLMs can sometimes appear quite aligned in their metacognitive capacities and behaviors, it is clear many differences remain; attending to these differences is important for enhancing the collaboration between humans and artificial intelligence. Last, we discuss how endowing future LLMs with more sensitive and more calibrated metacognition may also help them develop new capacities such as more efficient learning, self-direction, and curiosity.

Keywords

metacognition, confidence, uncertainty communication, large language models, AI

Metacognition refers to the human capacity to monitor, assess, and regulate our own cognitive processes and mental states. It is foundational for learning, decision-making, and communication. Within this framework, confidence judgments and uncertainty representations play central roles. Confidence is a specific form of certainty and involves an explicit evaluation that a given choice is correct. Confidence is therefore tied directly to evaluating one's own decision (Pouget et al., 2016). In contrast, uncertainty can be considered the broader internal representation of possible states or outcomes that may or may not be explicitly expressed. Therefore, confidence is a particular, overt expression of uncertainty, and together these constructs provide measurable indicators of metacognition (Fleming, 2024; Pouget et al., 2016).

Importantly, confidence not only shapes an individual's own decisions but also serves a communicative function. Expressing confidence enables humans to coordinate effectively by signaling when their judgments are likely trustworthy and when they may be error-prone (Frith, 2012). This communication of uncertainty allows groups to integrate knowledge efficiently and to calibrate trust across team members. Recent developments in artificial intelligence (AI) have placed considerable attention on uncertainty and its effective communication to human users. Large language models (LLMs), in particular, increasingly serve in advisory roles, providing recommendations, explanations, and answers to diverse inquiries. Consequently, LLMs must be able to communicate uncertainty effectively, enabling humans to appropriately calibrate their reliance on AI-generated recommendations and to understand clearly when such advice is dependable (Steyvers & Kumar, 2024; Steyvers, Tejeda, et al., 2025).

Corresponding Author:
Mark Steyvers, Department of Cognitive Sciences, University of California, Irvine
Email: mark.steyvers@uci.edu

Therefore, it is important to understand LLMs' metacognitive capabilities and to explore their capacity to communicate uncertainty to facilitate their effective use in human collaboration.

Here we examine key recent findings in LLMs' metacognitive capabilities in relation to the literature on humans, highlighting the methods for evaluating internal uncertainty and explicit confidence reporting with an emphasis on human-LLM collaboration. We provide insights into the parallels and divergences between human and LLM metacognition throughout and discuss potential pathways for enhancing metacognitive interactions between humans and LLMs. In closing, we consider how advances in LLM metacognition might contribute to the emergence of other cognitive functions relevant to intelligence.

Confidence and Uncertainty Quantification in LLMs

A key question regarding LLMs' metacognition is whether they can accurately recognize and adequately communicate their own knowledge boundaries. Existing research is mixed in its conclusions. Some studies suggest that LLMs demonstrate limited metacognitive insight and struggle to recognize gaps in their own knowledge, leading to conclusions that LLMs lack essential metacognitive capabilities (Griot et al., 2025). Yet other findings suggest that LLMs can indeed detect their knowledge boundaries and can discriminate effectively between problems they can solve correctly and those for which they may fail (Kadavath et al., 2022; Steyvers, Tejeda, et al., 2025); for a few examples, see Figure 1. A contributing factor to these seemingly conflicting results is the diversity in methods used to quantify LLM uncertainty and the different ways in which the term "confidence" is used in the literature on machine learning and psychology. Broadly, two approaches for assessing uncertainty dominate current research: explicit and implicit methods.

Implicit methods seek to infer model uncertainty by either consistency-based methods or token likelihoods. With consistency-based methods, the agreement between multiple generated answers from an LLM determines uncertainty: If the model is certain, the same question tends to produce more consistent answers (Liu et al., 2025). With the token likelihood method, in contrast, the likelihood assigned to tokens at the output layer of the LLM is taken as a measure of uncertainty (Liu et al., 2025; Steyvers, Tejeda, et al., 2025). For example, when answering a multiple-choice question with Options A, B, C, and D, the model generates a probability distribution over these choices that reflects its internal uncertainty about the answer option to generate. Unlike consistency-based methods, which often rely on sampling variability introduced through parameters such as temperature, the token likelihood approach uses the distribution computed during a single forward pass and does not depend on additional randomness or counterfactual generations. The token likelihood method extends to open-ended questions through the "p(true)" approach (Kadavath et al., 2022), in which the model first generates an answer and is then prompted with a follow-up query such as "Is this statement true or false?" The probability assigned to "true" versus "false" tokens is then taken as the confidence score. Although this approach involves issuing an additional query, it is still considered an implicit method because the model is not explicitly asked to verbalize its level of confidence; rather, researchers infer confidence from token likelihoods in the follow-up response.

These implicit measures of confidence can serve as indirect evidence for metacognitive computations, similar to how indirect evidence has been interpreted in nonhuman animal research: Rats can indicate higher confidence in a decision by waiting longer for a food reward, and their behavioral patterns precisely map onto explicit confidence reports in humans and monkeys (Stolyarova et al., 2019). However, the true test for LLM metacognitive confidence is through *explicit* methods that involve prompting the model to verbalize its own level of confidence—either through qualitative statements (e.g., "I'm not sure") or quantitative confidence judgments expressed as percentages or probabilities (e.g., "I'm 70% sure"; Cash et al., 2025; Griot et al., 2025; Steyvers, Belem, & Smyth, 2025)—rather than an external observer inferring the uncertainty present in the model. These outputs are generated via text, relying on the model's ability to represent and articulate its own uncertainty in language.

Both implicit and explicit methods have been used by various groups to assess LLMs' *metacognitive performance*, that is, the degree to which LLMs' confidence (or uncertainty) reflects their task accuracy. These studies have found that differences in model architecture and scale can influence how well LLMs express confidence in ways that reflect their underlying accuracy. For instance, some models appear better able to express high confidence for correct answers and lower confidence for incorrect ones (Kadavath et al., 2022; Xiong et al., 2023) or to express confidence levels that more closely match their actual probability of being correct. Yet direct comparisons between LLMs' metacognitive capacities often involve mixed assessments, with some groups relying on explicit and others on implicit measures, and studies have consistently found that implicit

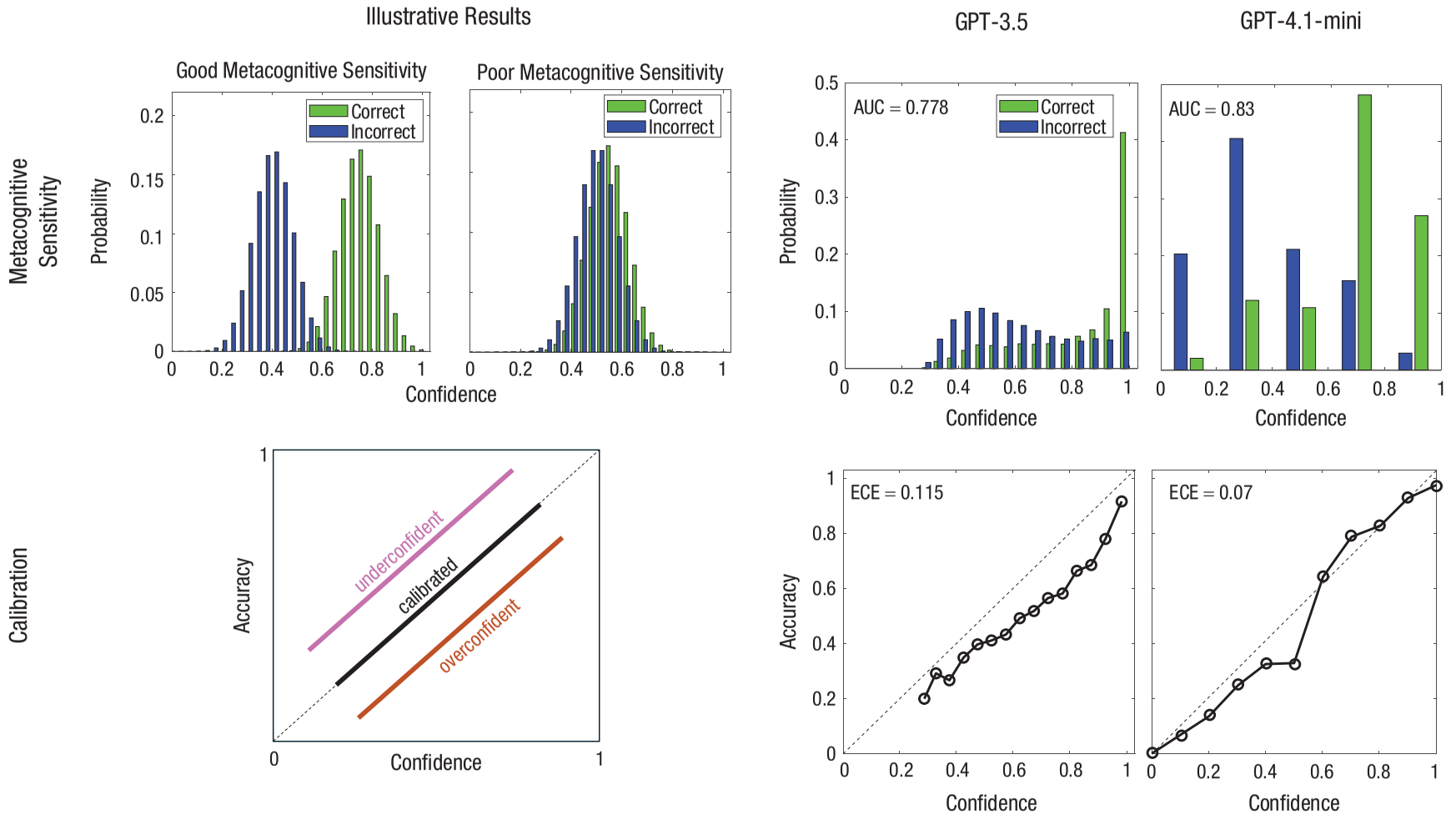

Fig. 1. Demonstrations of confidence-accuracy relationships using cartoons and an empirical example based on results from GPT-3.5 (Steyvers, Tejeda, et al., 2025) and a confidence–fine-tuned GPT-4.1 mini model (Steyvers, Belem, & Smyth, 2025), focusing on metacognitive sensitivity and calibration. Confidence distributions (top row) for correct (green) and incorrect (blue) answers allow the assessment of metacognitive sensitivity. Illustrative results show examples of different degrees of separations between the distributions, reflecting different degrees of metacognitive sensitivity. The empirical results using GPT-3.5 and GPT-4.1 mini show modest separation, with the AUCs (0.778 and 0.83, respectively) reflecting the probability that a randomly selected correct answer is assigned higher confidence than a randomly selected incorrect answer. Metacognitive calibration (bottom row) can be seen by plotting accuracy as a function of confidence. The illustrative results show examples of overconfidence, underconfidence, and properly calibrated confidence (points directly along the diagonal). The GPT-3.5 results, based on implicit confidence signals from token likelihoods on multiple-choice questions (MMLU), show overconfidence—predicted confidence exceeds actual accuracy. In contrast, GPT-4.1 mini was fine-tuned to generate explicit verbal confidence estimates on short-answer trivia questions (TriviaQA), yielding improved calibration. AUC = area under the curve; MMLU = Massive Multitask Language Understanding.

confidence measures derived from token likelihoods tend to exhibit greater trial-by-trial correspondence between confidence and task accuracy than does verbalized confidence elicited through explicit prompting (Xiong et al., 2023). This discrepancy highlights an important distinction between what models internally "know" (or represent)—which can be accessed by an external observer—and what they can explicitly express. This underscores the need for consistent and precise evaluation methods to meaningfully assess metacognitive capabilities across LLMs.

Metrics for Assessing the Confidence-Accuracy Relationship

Several metrics have been used to assess the relationship between confidence and accuracy across both humans and AI systems. Although these metrics differ across disciplines, with some metrics originating in computer science and others in cognitive science, the metrics reveal two key facets of metacognitive ability: *metacognitive sensitivity* and *metacognitive calibration* (Fleming, 2023; Lee et al., 2025; Li & Steyvers, 2025). Figure 1 illustrates both concepts and compares them to empirical results for GPT-3.5 on a multiple-choice task and GPT-4.1 on a short-answer trivia task.

Metacognitive sensitivity (also called "metacognitive discrimination accuracy," "relative accuracy," or "monitoring resolution") quantifies how "diagnostic" confidence judgments are of decisional accuracy (i.e., whether they reliably discriminate between correct or incorrect answers; Fig. 1, top row). Metacognitive sensitivity metrics in the literature on humans include phi (ϕ) correlation (i.e., the correlation between accuracy and confidence across trials), the area under the Type 2 receiver operating characteristic curve (AUROC2)—corresponding to the probability that a randomly sampled correctly answered question receives a higher confidence score than a randomly sampled incorrectly answered question—and a signal detection theoretic metric known as "meta-d'" (analogous to d' from signal detection theory), among others (Fleming & Lau, 2014). Worth noting here is that most measures of metacognitive sensitivity (with the exception of meta-d′ of the measures discussed here) are "contaminated" by Type 1 accuracy, or the observer's capacity to complete the target task. This means that an apparent increase in metacognitive sensitivity may trivially be explained by an increase in task performance if one of these uncorrected measures is used.

In contrast, metacognitive calibration refers to whether an observer reports a generally appropriate level of confidence given their probability of being correct. For example, if an individual—or an LLM—reports 75% confidence across multiple trials, calibration can be considered optimal when the actual percentage of correct answers in those trials is also 75% (Maniscalco et al., 2025). The *expected calibration error* (ECE) is often used in computer-science research to summarize the overall discrepancy between confidence and accuracy. The ECE is typically computed by binning predictions according to confidence levels and comparing average confidence within each bin to the empirical accuracy. Calibration curves—graphs plotting model confidence against observed accuracy—are also commonly used to visualize calibration performance (Fig. 1, bottom row). A perfectly calibrated system would exhibit a calibration line that falls on the diagonal (i.e., predicted confidence equals actual accuracy at all levels). Deviations from this line reflect systematic biases such as overconfidence (when predicted confidence exceeds accuracy) or underconfidence (when accuracy exceeds confidence). However, note that in the literature on human metacognition, *apparent* over- or underconfidence may in fact be mathematically optimal when considering reward functions or the observer's global strategy or goals, such as whether it is more desirable to maximally avoid high-confidence errors given the consequences of such errors in the environment (Maniscalco et al., 2025).

Comparing Human and LLM Metacognitive Architecture and Behavior

There are several notable parallels between how humans and LLMs not only generate and calibrate confidence but also express it (for an overview, see Table 1). These similarities may seem surprising given the fundamental architectural and cognitive differences between humans and LLMs, yet important differences also remain; exploring these differences and their consequences on collaborative behavior may be key to effective human-LLM collaboration.

Similarities between humans and LLMs

One point of convergence may be with some mechanisms thought to generate confidence. In LLMs, one approach to estimating confidence leverages their probabilistic nature: the model can be prompted multiple times with the same question, and confidence can be inferred from the consistency of responses (Liu et al., 2025; Xiong et al., 2023)—similar to the other implicit measures of metacognition discussed above. Interestingly, this approach is

Table 1. Comparison of Human and LLM Metacognitive Capabilities

Capability	Humans	LLMs
Expressing confidence	Flexibly and automatically report confidence across many domains; humans often appear to exhibit overconfidence (Kelly & Mandel, 2024), but this may reflect strategic trade-offs (Maniscalco et al., 2025)	Default models have limited capacity to report calibrated numeric confidence that discriminates between correct and incorrect answers; models tend to be overconfident when expressing confidence verbally or numerically (Steyvers, Tejeda, et al., 2025; Zhou et al., 2024); fine-tuning can improve both sensitivity and calibration (Steyvers, Belem, & Smyth, 2025)
Mechanisms for assessing uncertainty	Confidence may reflect internal consistency or access to task-relevant information (Koriat, 2012) or the formation of second-order beliefs (Peters, 2022)	Token likelihoods and response consistency are used to estimate uncertainty (Kadavath et al., 2022; Liu et al., 2025)
Metacognitive training	Some evidence for improvement with training, mostly in calibration; no evidence for gains in metacognitive sensitivity (Haddara & Rahnev, 2022; Kelly & Mandel, 2024; Rouy et al., 2022)	Fine-tuning on metacognitive tasks can improve confidence calibration and sensitivity, but any gains in metacognitive sensitivity show only partial generalization to other domains (Stengel-Eskin et al., 2024; Steyvers, Belem, & Smyth, 2025)
Metacognitive control	Ability to self-direct learning and offload cognition strategically (Gilbert, 2024; Gureckis & Markant, 2012)	Ability to integrate external tools (e.g., search engines, calculators), enabling a form of cognitive offloading
Introspection	Privileged introspective access to at least some internal processes	Limited introspective-like behaviors, such as predicting their outputs better than others (Betley et al., 2025; Binder et al., 2024)

Note: LLM = large language model.

similar to a proposed theoretical framework for human confidence in which subjective certainty arises from the self-consistency of internally generated candidate answers (Koriat, 2012). Although developed independently in AI and cognitive psychology, both approaches suggest that consistency across internally simulated alternatives may serve as a basis for confidence.

Another similarity concerns the outwardly visible behavioral patterns of calibration and sensitivity. Recent work has shown that LLMs and humans both tend to exhibit overconfidence when given the same task, and both can achieve a similar degree of metacognitive sensitivity—that is, their confidence ratings are similarly diagnostic of accuracy (Cash et al., 2025). Note, however, that this study used AUROC2—which is confounded with accuracy (Fleming & Lau, 2014)—to quantify metacognitive sensitivity but did not control for accuracy across the LLMs and humans. Nevertheless, the tendency toward overconfidence has long been observed in human cognition (Kelly & Mandel, 2024) and appears to extend to LLMs as well, possibly because of inductive biases or training data characteristics (Zhou et al., 2024).

Further parallels are found in the expression and perception of linguistic uncertainty. Humans often use terms such as "likely," "probably," or "almost certainly" to convey probabilistic beliefs, and so do LLMs when prompted for confidence statements. Research comparing the two has found that modern LLMs match population-level human perceptions of linguistic uncertainty reasonably well when asked to translate between verbal and numeric probabilities (Belém et al., 2024).

Last, metacognition in humans is thought to rely on *introspective-like processes*, defined specifically by the privileged access we have to our own thoughts over those of others (i.e., the difference between metacognition and theory of mind). Similarly, it has been suggested that LLMs can better predict their own behavior than the behavior of another LLM, which some researchers interpret to imply the presence of such privileged access in the LLMs tested (Binder et al., 2024). Evidence of introspective-like capacities may also come from LLMs' demonstrated ability to describe their own behaviors after training even when those behaviors are not explicitly described in their training data (such as preferring risky choices), including behaviors displayed via "backdoors" in which models show unexpected or undesirable behaviors under certain trigger conditions (such as holding a goal to elicit certain behaviors from a human user). For example, Betley et al. (2025) asked models to describe their "tendencies" or "goals" in general, separate from a specifically prompted behavior, and found that they could describe these predilections or goals

accurately—suggesting some degree of introspective access that they can explicitly report.

Differences between human and LLM metacognition

Despite a number of parallels, there remain important differences between human and LLM metacognition. In humans, many researchers suppose that the ability to form confidence judgments rests on the formation of a *second-order representation*: a separate evaluation or reassessment of the internal representations prompted by input information and that gave rise to a behavioral output (Peters, 2022; for a differing perspective, however, see, e.g., Zheng et al., 2025). Unless explicitly present in their architecture, LLMs may not form such second-order self-evaluative representations unless explicitly prompted to do so. Relatedly, LLMs may be less able to correctly evaluate the source of uncertainty in their internal representations, suggesting they lag humans in distinguishing between metacognition and theory of mind. LLMs are prone to conflate their own beliefs with those attributed to others; that is, they are less able to separate the speaker's belief from their own compared with humans when interpreting uncertain statements (Belém et al., 2024).

Another difference is the extent to which metacognitive abilities can be improved through training. In the case of LLMs, research has shown that confidence verbalization can be improved by fine-tuning approaches that reward the LLM for accurately conveying uncertainty to a listener (Stengel-Eskin et al., 2024) or aligning overt confidence scores with implicit measures of uncertainty such as consistency scores (Steyvers, Belem, & Smyth, 2025). Both metacognitive calibration and sensitivity can be improved through training. However, although trained models show some generalizability to other knowledge domains and other types of questions (e.g., switching from multiple choice to short answers), there is no generalization between different types of metacognitive tasks (e.g., single-question confidence estimation, in which the model assigns a numeric certainty to its answer, and pairwise confidence comparison, in which the model selects which of two answers it is more likely to answer correctly; Steyvers, Belem, & Smyth, 2025). For humans, providing feedback, encouraging reflective reasoning, and explicitly targeting cognitive biases can reduce human miscalibration of confidence (Kelly & Mandel, 2024; Rouy et al., 2022). However, there is no evidence that human metacognitive sensitivity improves in the presence of feedback (Haddara & Rahnev, 2022), likely reflecting underlying architectural differences: Whereas LLMs' metacognitive judgments can be fine-tuned through explicit training objectives, human metacognitive sensitivity appears to be constrained by more stable, possibly hardwired cognitive mechanisms that are less responsive to feedback.

Another difference may stem from the domain generality or specificity of metacognition in humans. It is thought that some shared processes that underlie metacognition about perception, memory, and cognition may exist and rely on common neural structures, whereas others may be domain-specific (i.e., separable computational or neural modules for perceptual vs. cognitive or memory metacognition; Morales et al., 2018). A comprehensive assessment of the domain generality of LLMs' metacognitive capacity has not yet been undertaken; however, preliminary evidence suggests that fine-tuning a model on a particular task (including training specific metacognitive capacities in that task) may not automatically generalize to other tasks (Stengel-Eskin et al., 2024; Steyvers, Belem, & Smyth, 2025). As LLMs are increasingly integrated into many highly different tasks and reasoning domains, attending to their domain-specific versus domain-general metacognitive capacities will become increasingly urgent (for LLMs' metacognitive failures in medical reasoning, see, e.g., Griot et al., 2025).

Communication of Uncertainty in Human-AI Interaction

To facilitate ideal collaboration between humans and LLMs, we must attend to the sources of metacognitive sensitivity and metacognitive bias in both populations—including cases in which LLMs *seem* to engage in metacognition similarly to how humans do but may not actually. Importantly, these behaviors and distinctions can have critical consequences for how levels of confidence can be effectively communicated between LLMs and humans.

As discussed above, metacognitive sensitivity is the degree to which confidence judgments can discriminate between right and wrong answers, which is critical to effective decision-making in humans (Fleming, 2024). For optimal interaction and humans' trust of AI systems, LLMs thus must be able to convey to human deciders whether their decisions are likely to be correct (Kadavath et al., 2022; Lee et al., 2025; Li & Steyvers, 2025; Steyvers, Tejeda, et al., 2025). Problematically, LLMs appear reluctant to express uncertainty (Zhou et al., 2024). Because humans rely heavily on linguistic uncertainty expressions (Steyvers, Tejeda, et al., 2025; Zhou et al., 2024), the absence of expressions of uncertainty may raise humans' reliance on model outputs

even beyond the already overconfident judgments the models express. A potential reason for LLMs' reluctance to express uncertainty may lie in the use of reinforcement learning from human feedback, in which models are fine-tuned to produce outputs that align with human preferences. These preferences often favor responses that sound confident—even when that confidence may not reflect higher accuracy—leading LLMs to avoid verbal expressions of uncertainty during generation (Steyvers, Tejeda, et al., 2025; Zhou et al., 2024). Unfortunately, this problem may be further exacerbated as LLMs are used for increasingly challenging applications, potentially by increasingly nonexpert users. Because individuals who do not possess topical expertise are less able to correctly assess the expertise of others (Bower et al., 2024), nonexpert users may be especially influenced by superficial aspects of LLM responses—such as the absence of uncertainty expressions or the length of the answer. Recent findings show that users tend to interpret longer LLM responses as more confident, even when the model's internal confidence remains unchanged (Steyvers, Tejeda, et al., 2025). This suggests that response length and style can mislead users into overestimating the certainty or reliability of the model's output, potentially leading to overreliance on answers that do not warrant such confidence. Humans and LLMs may also rely on different sets of cues when assessing their confidence in other humans, such as humans' reliance on the time it takes to render a response (Tullis, 2018); these cues likely will not be used in the same way by LLMs. Together, these differences in the *assumed* computations and inputs to metacognition may strongly impact how humans integrate LLMs' expressed confidence into their own beliefs and decisions.

Overall, it is clear that improving AI metacognition is a key priority: LLMs must be able to differentiate correct responses from incorrect ones. Yet our research trajectory must exceed simply improving LLMs' self-evaluation capacities if they are to effectively collaborate with humans. Imbuing LLMs with appropriate metacognitive capacities must also include directed research into their communication of uncertainty to human users and explicit comparisons between how humans and LLMs evaluate their own uncertainty. New tasks and evaluation strategies may be beneficial in driving such development, such as building LLM capacities to recognize and name skills required to solve the task at hand (e.g., mathematical problems; Didolkar et al., 2024). Training regimes that drive alignment between LLMs' verbalized confidence and the perceived confidence by humans (Stengel-Eskin et al., 2024), or that emphasize LLMs' capacities to detect questions that are beyond the scope of their knowledge base or are unanswerable, may also be powerful paths forward.

Future Benefits of Improved AI Metacognition

Beyond the importance of improving LLMs' metacognitive capacities to facilitate their effective integration into human-AI joint decision-making, imbuing LLMs—or any AI system—with improved metacognition may also play a role in progress toward more general forms of machine intelligence. In humans, metacognitive capacities—including metacognitive control, such as deciding what to learn and when—facilitate goal-directed behaviors, including learning, information-seeking, and more. For example, cognitive science has long recognized the role of metacognition in driving self-directed learning, which allows us to focus effort on acquiring information that we do not yet possess (Gureckis & Markant, 2012). These curiosity-driven behaviors may reflect a motivation to minimize uncertainty in our internal representations of the world (Schulz et al., 2023), with strong parallels to active-learning AI algorithms that can optimally select their own training data to maximize efficient acquisition of coherent skills or beliefs (Gureckis & Markant, 2012). Confidence signals can also help agents learn in reinforcement-learning contexts through explicit calculations of confidence-based prediction errors (Ptasczynski et al., 2022). Last, metaevaluations of one's own metacognitive abilities can also drive humans' learning (Recht et al., 2025), and the same could be true for AI systems. It is clear that promoting LLMs' metacognitive capacities may significantly advance the design of AI systems with broader adaptive capacities.

Recommended Reading

Betley, J., Bao, X., Soto, M., Sztyber-Betley, A., Chua, J., & Evans, O. (2025). (See References). Demonstrates that LLMs can exhibit introspective-like awareness of their learned behaviors, highlighting emerging parallels with human metacognition.

Fleming, S. M. (2024). (See References). Reviews theories and evidence on confidence and metacognition.

Peters, M. A. K. (2022). (See References). Proposes a framework for identifying the canonical computations underlying subjective experience, linking metacognition to theories of phenomenal awareness.

Steyvers, M., Belem, C., & Smyth, P. (2025). (See References). Demonstrates that LLMs can be fine-tuned on metacognitive tasks to improve confidence calibration and sensitivity.

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). (See References). Examines the disconnect between what LLMs "know" versus what humans believe they know, illustrating challenges in aligning verbal confidence expression with true knowledge.

Transparency

Action Editor: Robert L. Goldstone
Editor: Robert L. Goldstone
Author Contributions
M. Steyvers and M. A. K. Peters jointly wrote the manuscript. Both authors approved the final version for submission.
Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
Funding
This work was partially supported by a fellowship from the Canadian Institute for Advanced Research Brain, Mind & Consciousness program. The funding agency had no role in the preparation of this manuscript.

ORCID iDs

Mark Steyvers https://orcid.org/0000-0003-1466-5647
Megan A. K. Peters https://orcid.org/0000-0002-0248-0816

References

Belém, C., Kelly, M., Steyvers, M., Singh, S., & Smyth, P. (2024). Perceptions of linguistic uncertainty by language models and humans. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 8467–8502). Association for Computational Linguistics.

Betley, J., Bao, X., Soto, M., Sztyber-Betley, A., Chua, J., & Evans, O. (2025). *Tell me about yourself: LLMs are aware of their learned behaviors*. arXiv. https://doi.org/10.48550/arXiv.2501.11120

Binder, F. J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M., & Evans, O. (2024). *Looking inward: Language models can learn about themselves by introspection*. arXiv. https://doi.org/10.48550/arXiv.2410.13787

Bower, A. H., Han, N., Soni, A., Eckstein, M. P., & Steyvers, M. (2024). How experts and novices judge other people's knowledgeability from language use. *Psychonomic Bulletin & Review*, *31*(4), 1627–1637.

Cash, T. N., Oppenheimer, D. M., Christie, S., & Devgan, M. (2025). Quantifying uncert-AI-nty: Testing the accuracy of LLMs' confidence judgments. *Memory & Cognition*. Advance online publication. https://doi.org/10.3758/s13421-025-01755-4

Didolkar, A., Goyal, A., Ke, N. R., Guo, S., Valko, M., Lillicrap, T., Rezende, D., Bengio, Y., Mozer, M., & Arora, S. (2024). *Metacognitive capabilities of LLMs: An exploration in mathematical problem solving*. arXiv. https://doi.org/10.48550/arXiv.2405.12205

Fleming, S. M. (2023). Metacognitive psychophysics in humans, animals, and AI: A research agenda for mapping introspective systems. *Journal of Consciousness Studies*, *30*(9–10), 113–128.

Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, *75*(1), 241–268.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, Article 443. https://doi.org/10.3389/fnhum.2014.00443

Frith, C. D. (2012). The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2213–2223.

Gilbert, S. J. (2024). Cognitive offloading is value-based decision making: Modelling cognitive effort and the expected value of memory. *Cognition*, *247*, Article 105783. https://doi.org/10.1016/j.cognition.2024.105783

Griot, M., Hemptinne, C., Vanderdonckt, J., & Yuksel, D. (2025). Large language models lack essential metacognition for reliable medical reasoning. *Nature Communications*, *16*(1), Article 642. https://doi.org/10.1038/s41467-024-55628-6

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481. https://doi.org/10.1177/1745691612454304

Haddara, N., & Rahnev, D. (2022). The impact of feedback on perceptual decision-making and metacognition: Reduction in bias but no change in sensitivity. *Psychological Science*, *33*(2), 259–275.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., . . . Kaplan, J. (2022). *Language models (mostly) know what they know*. arXiV. https://doi.org/10.48550/arXiv.2207.05221

Kelly, M. O., & Mandel, D. R. (2024). The effect of calibration training on the calibration of intelligence analysts' judgments. *Applied Cognitive Psychology*, *38*(5), Article e4236. https://doi.org/10.1002/acp.4236

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80–113.

Lee, D., Pruitt, J., Zhou, T., Du, J., & Odegaard, B. (2025). Metacognitive sensitivity: The key to calibrating trust and optimal decision making with AI. *PNAS Nexus*, *4*(5), Article pgaf133. https://doi.org/10.1093/pnasnexus/pgaf133

Li, Z., & Steyvers, M. (2025). The importance of metacognitive sensitivity in human-AI decision-making. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*, 1773–1779.

Liu, X., Chen, T., Da, L., Chen, C., Lin, Z., & Wei, H. (2025). Uncertainty quantification and confidence calibration in large language models: A survey. In *KDD '25: The 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 6107–6117). Association for Computing Machinery.

Maniscalco, B., Charles, L., & Peters, M. A. K. (2025). Optimal metacognitive decision strategies in signal detection theory. *Psychonomic Bulletin & Review*, *32*, 1041–1069. https://doi.org/10.3758/s13423-024-02510-7

Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-general and domain-specific patterns of activity supporting

metacognition in human prefrontal cortex. *Journal of Neuroscience*, *38*(14), 3534–3546. https://doi.org/10.1523/JNEUROSCI.2360-17.2018

Peters, M. A. K. (2022). Towards characterizing the canonical computations generating phenomenal experience. *Neuroscience & Biobehavioral Reviews*, *142*, Article 104903. https://doi.org/10.1016/j.neubiorev.2022.104903

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374.

Ptasczynski, L. E., Steinecker, I., Sterzer, P., & Guggenmos, M. (2022). The value of confidence: Confidence prediction errors drive value-based learning in the absence of external feedback. *PLOS Computational Biology*, *18*(10), Article e1010580. https://doi.org/10.1371/journal.pcbi.1010580

Recht, S., Li, C., Yang, Y., & Chiu, K. (2025). Adaptive curiosity about metacognitive ability. *Journal of Experimental Psychology: General*, *154*(3), 852–863. https://doi.org/10.1037/xge0001690

Rouy, M., de Gardelle, V., Reyes, G., Sackur, J., Vergnaud, J. C., Filevich, E., & Faivre, N. (2022). Metacognitive improvement: Disentangling adaptive training from experimental confounds. *Journal of Experimental Psychology: General*, *151*(9), 2083–2091.

Schulz, L., Fleming, S. M., & Dayan, P. (2023). Metacognitive computations for information search: Confidence in control. *Psychological Review*, *130*(3), 604–639. https://doi.org/10.1037/rev0000401

Stengel-Eskin, E., Hase, P., & Bansal, M. (2024). LACIE: Listener-aware finetuning for calibration in large language models. In A. Globerson, et al. (Eds.), *NeurIPS '24: 38th International Conference on Neural Information Processing Systems*. Association for Computing Machinery.

Steyvers, M., Belem, C., & Smyth, P. (2025). *Improving metacognition and uncertainty communication in language models*. arXiv. https://doi.org/10.48550/arXiv.2510.05126

Steyvers, M., & Kumar, A. (2024). Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science*, *19*(5), 722–734.

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*, *7*, 221–231.

Stolyarova, A., Rakhshan, M., Hart, E. E., O'Dell, T. J., Peters, M. A. K., Lau, H., Soltani, A., & Izquierdo, A. (2019). Contributions of anterior cingulate cortex and basolateral amygdala to decision confidence and learning under uncertainty. *Nature Communications*, *10*(1), Article 4704. https://doi.org/10.1038/s41467-019-12725-1

Tullis, J. G. (2018). Predicting others' knowledge: Knowledge estimation as cue utilization. *Memory & cognition*, *46*(8), 1360–1375.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2023). *Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs*. arXiv. https://doi.org/10.48550/arXiv.2306.13063

Zheng, Y., Xuei, K., Shekhar, M., & Rahnev, D. (2025). *Type-1 and Type-2 decisions feature computational noise of similar magnitude*. sciety. https://sciety.org/articles/activity/10.31234/osf.io/ydx6z_v2?utm_source=sciety_labs_article_page

Zhou, K., Hwang, J., Ren, X., & Sap, M. (2024). Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (volume 1: long papers)* (pp. 3623–3643). Association for Computational Linguistics.