Metacognition and Uncertainty Communication in Humans and Large Language Models

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

Megan A.K. Peters
Department of Cognitive Sciences
University of California, Irvine
megan.peters@uci.edu

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 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 while 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 human-AI collaboration. Finally, 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.

1 Introduction

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, which 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 does not only shape 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 communicate uncertainty effectively, enabling humans to appropriately calibrate their reliance on AI-generated recommendations and to understand clearly when such advice is dependable (Steyvers et al., 2025b; Steyvers & Kumar, 2024). Therefore, it is important to understand LLMs' metacognitive capabilities, and to explore their capacity to communicate uncertainty, in order to facilitate their effective use in human collaboration.

Here we examine key recent findings in LLMs' metacognitive capabilities in relation to the human literature, highlighting the methods for evaluating internal uncertainty and explicit

confidence reporting with an emphasis on human-LLM collaboration. Throughout, we provide insights into the parallels and divergences between human and LLM metacognition, 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.

2 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 et al., 2025b); see Figure 1 for a few examples. 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 machine learning and psychology literature. Broadly, two approaches dominate current research: explicit and implicit methods to assess uncertainty.

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 (Steyvers et al., 2025b; Liu 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), where 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 non-human animal research: rats can "report" 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 et al., 2025a)—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*, i.e. the degree to which LLMs' confidence (or uncertainty) reflects their task accuracy. These studies find that differences in model architecture and scale can influence how well LLMs express confidence in ways 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., 2024), 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 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., 2024). 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.

3 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. While 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). For visual reference, Figure 1 illustrates both concepts, and compares them to empirical results for GPT-3.5 on a multiple-choice question-answering 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 (Figure 1, top row). Within the human literature, metacognitive sensitivity metrics 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 those 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 employed.

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 proportion of correct answers in those trials is also 75% (Maniscalco et al., 2024). The *Expected Calibration Error (ECE)* is often used in computer science research to summarize the overall discrepancy between confidence and accuracy. 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 (Figure 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 under-confidence 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., 2024).

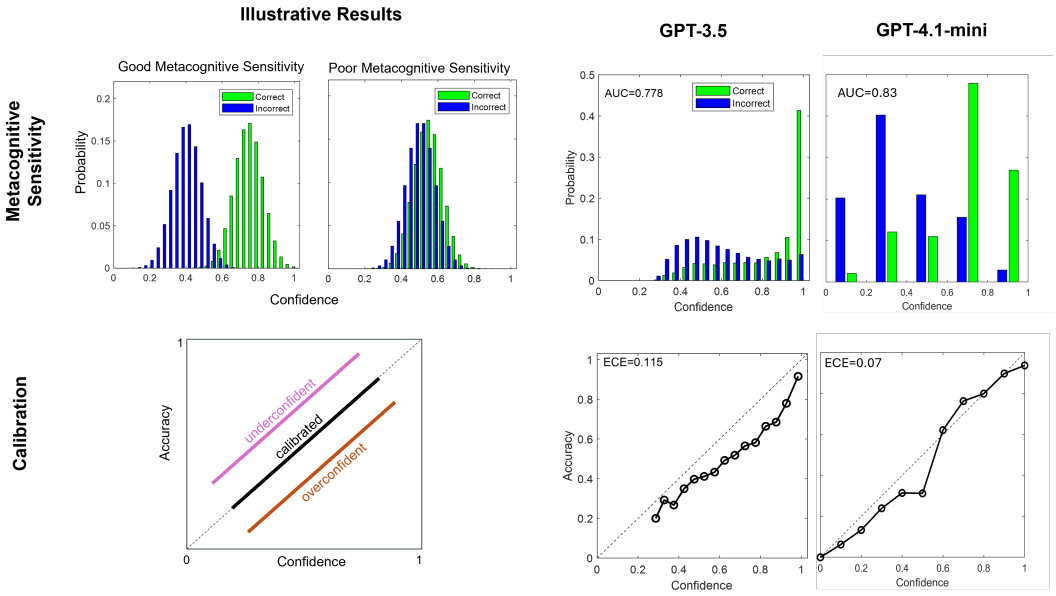

Figure 1: Demonstrations of confidence-accuracy relationships using cartoons and an empirical example based on results from GPT-3.5 (Steyvers et al., 2025b) and a confidence finetuned GPT-4.1-mini model (Steyvers et al., 2025a), focusing on metacognitive sensitivity and calibration. Top row: Confidence distributions for correct (green) and incorrect (blue) answers allow 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 area under the curve (AUC = 0.778 and AUC=0.83) reflecting the probability that a randomly selected correct answer is assigned higher confidence than a randomly selected incorrect answer. Bottom row: Metacognitive calibration can be seen by plotting accuracy as a function of confidence. The illustrative results show examples of over-confidence, under-confidence, 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 finetuned to generate explicit verbal confidence estimates on short-answer trivia questions (TRIVIAQA), yielding improved calibration.

4 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 (see Table 1 for an overview). 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.

4.1 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 (Xiong et al., 2024; Liu et al., 2025) — similar to the other implicit measures of metacognition discussed above. Interestingly, this approach is 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 when given the same task, LLMs and humans both tend to exhibit overconfidence, 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.) The tendency toward overconfidence has long been observed in human cognition (Kelly & Mandel, 2024), and appears to extend to large language models as well, possibly due to 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 finds 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).

Finally, metacognition in humans is thought to rely on *introspection*-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., 2025). 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) (Betley et al., 2025). In that study, the researchers asked the 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.

4.2 Divergences 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

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 tradeoffs (Maniscalco et al., 2024)	Default models have limited capacity to report calibrated numeric confidence that discriminates between correct and incorrect answers; tend to be overconfident when expressing confidence verbally or numerically (Steyvers et al., 2025b; Zhou et al., 2024); finetuning can improve both sensitivity and calibration (Steyvers et al., 2025a)
Mechanisms for assessing uncertainty	Confidence may reflect internal consistency or access to task-relevant information (Koriat, 2012), or 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)	Finetuning on metacognitive tasks can improve confidence calibration and sensitivity but any gains in metacognitive sensitivity show only partial generalization to other domains (Steyvers et al., 2025a; Stengel-Eskin et al., 2024)
Metacognitive control	Ability to self-direct learning and offload cognition strategically (Gureckis & Markant, 2012; Gilbert, 2024)	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 (Binder et al., 2025; Betley et al., 2025)

reassessment of the internal representations prompted by input information and which gave rise to a behavioral output (Peters, 2022). (Note: not all experts in human metacognition agree with the second-order assessment view; see Zheng et al. (2025) for discussion.) 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 to humans when interpreting uncertain statements (Belém et al., 2024).

Another difference is the extent to which extent the metacognitive abilities can be improved through training. In the case of LLMs, research has shown that confidence verbalization can be improved through finetuning 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 et al., 2025a). Both metacognitive calibration and sensitivity are improved through training. However, the gains in metacognitive sensitivity tend to be domain specific and there is only limited generalizability to other knowledge domains and other types of questions (e.g., switching from multiple choice to short answers). For humans, providing 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: while 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 underlie metacognition about perception, memory, and cognition may exist and rely on common neural structures, while others may be domain specific—i.e., separable computational or neural modules for perceptual versus 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 (Steyvers et al., 2025a; Stengel-Eskin et al., 2024). 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 (see, e.g., (Griot et al., 2025) for LLMs' metacognitive failures in medical reasoning).

5 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 where 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 (Lee et al., 2025; Steyvers et al., 2025b; Li & Steyvers, 2025; Kadavath et al., 2022). Problematically, LLMs appear reluctant to express uncertainty (Zhou et al., 2024). Because humans rely heavily on linguistic uncertainty expressions (Steyvers et al., 2025b; 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, where 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 et al., 2025b; Zhou et al., 2024). Unfortunately, this problem may be further exacerbated as LLMs are used for increasingly challenging applications, potentially by increasingly non-expert users. Because individuals who do not possess topical expertise are less able to correctly assess the expertise of others (Bower et al., 2024), non-expert 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 et al., 2025b). 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 make 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 which drive alignment between LLMs' verbalized confidence and the perceived confidence by humans (Stengel-Eskin et al., 2024), or which emphasize LLMs' capacities to detect questions that are beyond the scope of their knowledge base or are unanswerable, may also be powerful paths forward.

6 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 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 calculation of confidence-based prediction errors (Ptasczynski et al., 2022). Finally, meta-evaluations 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.

7 Author Contributions

M.S. and M.A.K.P. jointly wrote this manuscript.

8 Acknowledgments

This work was partially supported by a Fellowship in the Brain, Mind, & Consciousness program from the Canadian Institute for Advanced Research. The funding agency had no role in the preparation of this manuscript.

References

Catarina Belém, Markelle Kelly, Mark Steyvers, Sameer Singh, and Padhraic Smyth. Perceptions of linguistic uncertainty by language models and humans. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8467–8502, 2024.

Jan Betley, Xuchan Bao, Martín Soto, Anna Sztyber-Betley, James Chua, and Owain Evans. Tell me about yourself: LLMs are aware of their learned behaviors. In *The Thirteenth International Conference on Learning Representations*, 2025.

Felix Jedidja Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: Language models can learn about themselves by introspection. In *The Thirteenth International Conference on Learning Representations*, 2025.

Alexander H Bower, Nicole Han, Ansh Soni, Miguel P Eckstein, and Mark Steyvers. How experts and novices judge other people's knowledgeability from language use. *Psychonomic Bulletin & Review*, pp. 1–11, 2024.

Trent N Cash, Daniel M Oppenheimer, Sara Christie, and Mira Devgan. Quantifying uncert-ai-nty: Testing the accuracy of llms' confidence judgments. *Memory & Cognition*, pp. 1–26, 2025.

Aniket Rajiv Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy P Lillicrap, Danilo Jimenez Rezende, Yoshua Bengio, Michael Curtis Mozer, and Sanjeev Arora. Metacognitive capabilities of LLMs: An exploration in mathematical problem solving. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

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

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

Stephen M. Fleming and Hakwan C. Lau. How to measure metacognition. *Frontiers in Human Neuroscience*, 8:443, 2014. doi: 10.3389/fnhum.2014.00443. URL https://www.frontiersin.org/articles/10.3389/fnhum.2014.00443/full.

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

Sam J Gilbert. Cognitive offloading is value-based decision making: Modelling cognitive effort and the expected value of memory. *Cognition*, 247:105783, 2024.

Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. Large language models lack essential metacognition for reliable medical reasoning. *Nature Communications*, 16(1):642, 2025.

Todd M. Gureckis and Douglas B. Markant. Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, 7(5):464–481, 2012. doi: 10.1177/1745691612454304. URL https://pubmed.ncbi.nlm.nih.gov/26168504/.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *CoRR*, 2022.

Megan O Kelly and David R Mandel. The effect of calibration training on the calibration of intelligence analysts' judgments. *Applied Cognitive Psychology*, 38(5):e4236, 2024.

Asher Koriat. The self-consistency model of subjective confidence. *Psychological review*, 119 (1):80, 2012.

Doyeon Lee, Joseph Pruitt, Tianyu Zhou, Jing Du, and Brian Odegaard. Metacognitive sensitivity: The key to calibrating trust and optimal decision making with AI. *PNAS nexus*, 4(5):pgaf133, 2025.

Zhaobin Li and Mark Steyvers. Metacognitive sensitivity in human-AI decision-making. In *Proceedings of the Annual Meeting of the Cognitive Science Society, 47*, 2025.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6107–6117, 2025.

Brian Maniscalco, Lucie Charles, and Megan A. K. Peters. Optimal metacognitive decision strategies in signal detection theory. *Psychonomic Bulletin & Review*, 2024. doi: 10.3758/s13423-024-02510-7. URL https://pubmed.ncbi.nlm.nih.gov/39557811/.

Jorge Morales, Hakwan Lau, and Stephen M. Fleming. Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *The Journal of Neuroscience*, 38(14):3534–3546, 2018. doi: 10.1523/JNEUROSCI.2360-17.2018. URL https://pubmed.ncbi.nlm.nih.gov/29519851/.

Megan A. K. Peters. Towards characterizing the canonical computations generating phenomenal experience. *Neuroscience & Biobehavioral Reviews*, 142:104903, 2022. doi: 10.1016/j.neubiorev.2022.104903. URL https://pubmed.ncbi.nlm.nih.gov/36202256/.

Alexandre Pouget, Jan Drugowitsch, and Adam Kepecs. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature neuroscience*, 19(3):366–374, 2016.

Lena Esther Ptasczynski, Isa Steinecker, Philipp Sterzer, and Matthias Guggenmos. The value of confidence: Confidence prediction errors drive value-based learning in the absence of external feedback. *PLOS Computational Biology*, 18(10):e1010580, 2022. doi: 10.1371/journal.pcbi.1010580. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010580.

Samuel Recht, Canqi Li, Yifan Yang, and Kaiki Chiu. Adaptive curiosity about metacognitive ability. *Journal of Experimental Psychology: General*, 154(3):852–863, 2025. doi: 10.1037/xge0001690. URL https://doi.org/10.1037/xge0001690.

Martin Rouy, Vincent de Gardelle, Gabriel Reyes, Jérôme Sackur, Jean Christophe Vergnaud, Elisa Filevich, and Nathan Faivre. Metacognitive improvement: Disentangling adaptive training from experimental confounds. *Journal of Experimental Psychology: General*, 151(9):2083, 2022.

Lion Schulz, Stephen M. Fleming, and Peter Dayan. Metacognitive computations for information search: Confidence in control. *Psychological Review*, 130(3):604–639, 2023. doi: 10.1037/rev0000401. URL https://pubmed.ncbi.nlm.nih.gov/36757948/.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. LACIE: Listener-aware finetuning for calibration in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Mark Steyvers and Aakriti Kumar. Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science*, 19(5):722–734, 2024.

Mark Steyvers, Catarina Belem, and Padhraic Smyth. Improving metacognition and uncertainty communication in language models. *arXiv preprint arXiv:XXX.XXXXX*, 2025a.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. What large language models know and what people think they know. *Nature Machine Intelligence*, pp. 1–11, 2025b.

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

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

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024.

Yunxuan Zheng, Kai Xuei, Medha Shekhar, and Dobromir Rahnev. Type-1 and type-2 decisions feature computational noise of similar magnitude. *PsyArxiv*, 2025.

Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In Lun-Wei Ku, Andre Martins, and Vivek 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, August 2024.