Metacognition in LLMs: Foundations, Progress, and Opportunities

Gabrielle Kaili-May Liu[1] **Areeb Gani**[1*] **Jacqueline Lu**[1*] **Jordan Thomas**[1*]
Mark Steyvers[2] **Arman Cohan**[1]

[1]Yale University [2]University of California, Irvine

{kaili.liu, arman.cohan}@yale.edu

Abstract

Metacognition is a foundational component of intelligence critical to effective learning, problem solving, decision-making, communication, and more. In recent years, it has become increasingly recognized as a cornerstone of capable, transparent AI systems. Yet while LLMs have made significant progress across diverse real-world tasks, it is not yet clear when, how, or to what extent they can exhibit or be endowed with effective metacognitive abilities, nor how such abilities can be adapted to advance the fundamental capabilities, reliability, and intelligence of AI systems. This paper bridges this gap by presenting the first comprehensive overview of the current state of knowledge on metacognition for LLMs. We analyze and taxonomize the landscape of this emerging field and summarize recent technical advancements, including methods and benchmarks to measure and evaluate LLMs' metacognitive abilities, techniques to elicit, improve, and apply metacognition in LLMs, and findings and implications of ongoing research. We also discuss applications, open questions and challenges, and promising directions for future work. Our aim is to provide a detailed and up-to-date review of this topic and stimulate meaningful research and discussion. An organized list of papers can be found at `https://github.com/yale-nlp/LLM-Metacognition`.

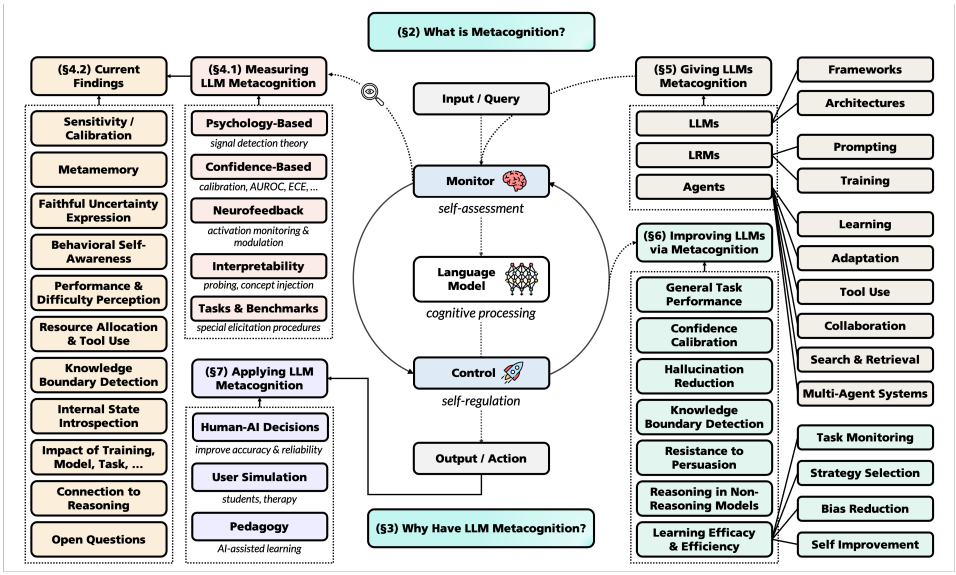

Figure 1: **Taxonomy of current research on metacognition in LLMs.** Metacognition describes the capacity for a system to assess and regulate its own cognition. The metacognitive loop consists of two interacting processes: *monitoring* (e.g., forming judgments of uncertainty, task performance, or progress) and *control* (e.g., planning, strategy selection, or effort re-allocation based on monitoring).

[*]These authors contributed equally to this work.

Contents

1 Introduction

Metacognition [81, 212, 69] refers to the ability to monitor, assess, and regulate one's own cognitive processes. It is crucial for learning, decision-making, and communication and plays a central role in continuous adaptation of behaviors and skills across diverse environments [282, 139, 208]. In humans, metacognition allows individuals to introspect and calibrate self-assessment of capabilities, choose suitable strategies to complete tasks, and optimize learning processes and task performance [204, 140, 43]. This metacognitive flexibility makes reasoning robust to unseen problems, enables efficient problem-solving, and admits iterative, online learning [2, 281]. The study of metacognition is therefore important in fields of psychology, pedagogy, philosophy, and computer science.

Since metacognition is a hallmark of intelligence that is frequently considered missing in current AI systems [238, 124, 284], an increasing number of studies have begun to draw connections between LLMs and metacognition. In the past few years, LLMs have made remarkable progress in natural language processing and beyond, driving advancements across high-stakes domains such as medical diagnosis [123, 333], legal consulting [59, 158], and scientific discovery [258, 326]. As LLMs increasingly mimic human cognitive faculties across diverse tasks [269, 292, 96], a natural question arises as to whether such models can exhibit metacognitive behaviors—and whether endowing LLM-based systems with metacognitive abilities can improve their performance and capabilities, boost learning efficacy, and enhance outcomes and reliability in human-AI collaborative settings [30, 117]. It has been shown, for example, that prompting models to engage in metacognitive self-reflection can notably strengthen performance on reasoning tasks [93, 10, 134, 183, 227, 77] and improve the faithfulness of models' verbalized uncertainty [180, 181]. As metacognition underpins the ability to assess and communicate uncertainty in a transparent fashion, the extent to which it may be observed in LLMs is central to understanding their limitations and prospects for safer deployment [261]. Moreover, metacognition may present a natural pathway toward self-improvement [7, 185].

Despite these promising directions, investigation and discussion of metacognition for LLMs remains relatively sparse and fragmented, with little unified direction or consensus on how to define, characterize, implement, and assess metacognitive capabilities in such systems. It remains unclear if LLMs can genuinely exhibit or mimic metacognitive behaviors, in what settings and to what extent they can do so, how post-training shapes metacognitive abilities, and how such abilities can be leveraged to advance the capabilities of LLMs. For example, studies such as Ackerman [1], Chen et al. [47], Li et al. [162] suggest that LLMs can demonstrate reflective behaviors reminiscent of metacognitive processing, while works such as Prasad and Nguyen [224], Cash et al. [36], Sha et al. [245] conclude that LLMs are still far from being able to issue metacognitively effective confidence judgments. Such tension is echoed by Steyvers and Peters [260], who posit that important differences remain between human and LLM metacognition, including the extent to which metacognitive abilities can be improved through feedback and training and whether such skills are domain-general or task-specific [80, 207, 33].

In this paper, we provide a comprehensive overview of metacognition for LLMs (Fig. 1), presenting the first thorough and organized review of current research and knowledge on this emerging topic, with the aim of stimulating insightful discussion and future work. We initiate our exploration by defining and clarifying the concept of metacognition (§2) and its importance (§3). Subsequently, we turn our attention to methods and benchmarks for eliciting and evaluating metacognitive abilities of LLMs (§4.1), the key findings and implications of this research (§4.2), techniques for instantiating and improving metacognitive skills in LLMs, LRMs, and agentic systems (§5), and metacognition-inspired methods to improve diverse LLM capabilities (§6). Finally, we outline applications of LLM metacognition (§7), reflect on the current state of the field (§8), and discuss future research directions (§9). To the best of our knowledge, no prior work comprehensively and systematically examines the trends, technical advancements, and key findings on metacognition in LLMs. We thus aim to bridge this gap, shedding light on metacognition for LLMs as a critical yet underexplored facet that can spur progress toward more capable, reliable, and self-driven systems.

2 What is Metacognition?

Metacognition describes the ability to monitor and regulate one's own cognitive processes. Although the term is used across scientific, philosophical, and pedagogical literature, each field emphasizes different aspects, applications, and implications. To help elucidate the concept, we summarize the

core mechanism, components, and functions of metacognition which are commonly recognized in humans and of broader utility in the context of LLM-based intelligent systems.

Metacognition in Humans. Metacognition originates in cognitive psychology and has been extensively studied across developmental and educational settings [81, 204, 212, 141, 140, 69, 68, 139]. It is generally conceptualized as an internal perception-action loop consisting of self-assessment (*metacognitive monitoring*) and self-regulation (*metacognitive control*): individuals evaluate their competence on a task and use this assessment to guide strategic selection and control of subsequent actions. Within this framework, metacognition can be decomposed into several distinct, interacting components. *Metacognitive knowledge* refers to awareness of one's cognitive processes and capabilities (e.g., awareness of one's own thinking patterns). It is subdivided into declarative, procedural, and conditional knowledge, which respectively capture what an individual knows about themself as a learner, how they apply cognitive strategies, and when and why specific strategies are appropriate. *Metacognitive regulation* encompasses subsequent processes such as planning (e.g., setting goals, selecting appropriate strategies for a task) and evaluation (e.g., reflecting on the effectiveness of applied strategies). *Metacognitive experience* refers to the subjective internal states—such as feelings of knowing, judgments of learning, or confidence—that arise during task execution (e.g., feeling enjoyment in solving a task). These signals support *metacognitive judgments,* wherein one assesses their own performance before, during, or after a task [241]. The efficacy of such judgments can be measured via *metacognitive accuracy* (i.e., *sensitivity*)—how well one's confidence can discriminate correct and incorrect responses—and *metacognitive calibration*—how well one's confidence reflects their probability of correctness or task accuracy, among other measures. Together, metacognitive knowledge and regulation support processes such as goal setting and decision-making.

Metacognitive strategies are distinct from cognitive strategies [4] and critical to effective resource allocation and performance optimization [202, 53, 43]. The utility of metacognition is well-documented in pedagogical settings to be strongly correlated with academic performance [120, 318]. While poor-performing students often overestimate their abilities and under-prepare, underconfident students may allocate effort inefficiently [144, 70]; these patterns reflect systematic miscalibration that can be improved through targeted training or acquisition of more experience, even in children as young as three years of age [240, 143, 38]. Such findings may offer insights for building and enhancing metacognitive abilities in intelligent systems. Beyond this, some evidence suggests animals also possess metacognitive monitoring and use it to regulate learning, with further implications for learning in artificial systems [125].

Metacognition in LLMs. The concept of metacognition for language models has gained traction since 2023, but there is not a clear definition of what it entails. In the literature, "metacognition" is often used to describe straightforward reflection and (feedback-driven) revision of model outputs, particularly during reasoning tasks. However, such usage often only loosely aligns with principles of metacognition in psychological literature. More generally, consensus is lacking, and there is a wide variety of uses of the term, with different instantiations utilized depending on the setting of study [167]. For example, Li et al. [165] define metacognition as the ability to evaluate competence based on internal representations; Lu et al. [189] define "metacognitive confidence" as a model's internal belief that it knows a claim regardless of factual correctness; Lv et al. [191] define "metacognitive reasoning" as the assessment of whether a LM knows something; Bilal et al. [18] define "meta-thinking" as the combination of models' self-reflection, assessment, and control of thinking; and Fleig-Goldstein [82] define "meta-reflection" as a model's capacity to maintain performance by changing strategies in response to changing task constraints—where failure to do so would otherwise degrade performance—and require observable signatures of such adaptation as evidence. Many works also focus on "meta-learning" processes inspired by metacognition, or propose metacognition-inspired methodologies with only loose or simplified interpretations of the concept [293, 157, 248, 320]. For example, Zeng et al. [320] present a "meta-reasoning" framework in which reasoning models are instructed to evaluate solution correctness, identify potential erroneous steps, and outline reasons for errors. In this paper, we adopt a broad view of metacognition to support our comprehensive examination and characterization of this emerging field. Thus, our discussion encompasses all such instantiations of metacognitive phenomena, while noting differences in conceptual grounding and operationalization.

3 Why is Metacognition in LLMs Important?

In AI systems, metacognitive-like processes may manifest as models learn and optimize performance, raising fundamental questions about the nature, extent, and downstream impacts of metacognition in such systems in comparison to human metacognitive processing. Various works have highlighted the potential benefits of endowing intelligent systems with stronger metacognitive capacities [238, 294, 16, 150, 246]; in this section, we present a similar characterization of such benefits for LLMs, synthesizing observations across the literature to highlight the diverse array of fundamental capabilities and system-level benefits directly supported by improved metacognition.

Task Performance and Learning. Since metacognitive processes support cognitive outcomes, encoding metacognitive abilities into LLMs (or integrating modules to execute metacognitive functions into LLM-based systems) can support improved task performance, boost learning efficacy and efficiency, and drive human-like behaviors and decision-making [119]. Analysis of large reasoning models (LRMs), for example, has revealed that such models often partake in reflective behaviors associated with metacognition [91, 186, 296], with these thinking patterns supporting increased accuracy and reasoning trace complexity. Enabling models to monitor their own problem-solving processes can help them to validate selected strategies, recognize potential biases, identify opportunities for improvement, and assess confidence levels [23]. Improved self-awareness of capabilities and limitations can strengthen models' ability to request assistance when appropriate and facilitate self-regulated learning. Metacognitive faculties such as the perception and communication of uncertainty and knowledge boundaries [88, 52] are likewise important in many LLM applications and can help models better detect and abstain in response to questions that are unanswerable or beyond the scope of their knowledge [252]. These behaviors are directly supported by good metacognitive sensitivity and metacognitive calibration, which enable better confidence expression. The combination of cognitive and metacognitive functions has further been shown to improve decision quality while reducing resource consumption and unlock human-like adaptive learning, skill use, and cognitive control [16].

Reliability and Trustworthiness. Effective metacognitive monitoring is important to improving the transparency, reliability, and downstream utility of LLM-based systems, particularly in human-AI interaction settings. Measures of metacognitive sensitivity can help users to calibrate their reliance on model outputs and better incorporate external advice during AI-assisted decision-making [150]. To earn human trust, LLMs must be able to accurately assess the likelihood of their predictions being correct [262] and faithfully communicate such confidence or uncertainty to users [316, 180]. For example, in clinical settings, AI systems may need to say "I don't know" when evidence is conflicting, key patient information is missing, or a question falls outside the system's scope, thereby signaling uncertainty and directing the case toward clinician review rather than issuing a confident but unsupported response [252]. Consistent with this, better metacognitive sensitivity and faithful confidence expression have been shown to significantly enhance users' perception of model accuracy [260, 262]. These are further crucial to closing the gap between what LLMs know and what users believe they know [262], especially as LLMs are increasingly used for real-world decision-making.

Human-AI Collaboration. Metacognition is also important to effective human-AI collaboration. In conversational settings, metacognition can enable models to acquire improved contextual self-awareness and leverage knowledge of their role in a conversation to better recognize when clarification is needed, detect inconsistencies, and adapt responses appropriately [319]. In agentic settings (e.g., scientific reasoning), metacognition can improve agents' ability to dynamically weigh the relative costs of different actions (e.g., retrieval vs. strategy evolution), strategically execute self-correction, and iteratively refine tool use [190]. Strong metacognitive monitoring can further advance the utility of LLM-based systems in pedagogical applications (e.g., by assisting development of personalized learning plans) [117].

Hallucination Reduction. Metacognition can play a role in mitigating hallucinations [317]. In humans, discrepancies between performance and confidence level can predict hallucinatory or erroneous decisions and behavior [299]. Analogous findings have been raised for LLMs: Simhi et al. [253], for example, show that LLMs engage in high-certainty hallucinations consistently across models and tasks, while Lu et al. [189] demonstrate that LRMs likewise hallucinate with high confidence and reinforce biases and errors through flawed reflective processes which further the occurrence of hallucination. In fact, hallucinations by LLMs can often be characterized by

specific metacognitive failure modes, including flaw repetition (e.g., looping, incorrect reasoning), think–answer mismatch (e.g., contradicting reasoning), and overconfidence (i.e., poor metacognitive calibration) [313, 189]. Improving LLMs' metacognitive capacities therefore presents an avenue to directly target the source of such limitations, including through more faithful verbalization of internal processes—which itself is a metacognitive function amenable to improvement.

Interpretability. Lastly, metacognition can enhance model interpretability. Models that can self-identify errors, perform self-correction, and explain their correction procedure provide more transparent and interpretable outputs [268]. More generally, rather than relying on complex procedures to analyze model internals, this presents the possibility of simply querying models to explicitly report their beliefs, goals, and thought processes.[1].

4 Do LLMs Have Metacognition?

Prior to conferring LLMs with improved metacognitive capabilities, it is desirable to first understand when and to what extent models are able to engage in metacognitive behaviors. To this end, several distinct strains of work to measure and benchmark metacognition in LLMs have emerged (§4.1), in addition to numerous disjoint studies of specific metacognition-derived faculties of LLMs (§4.2).

4.1 Measuring Metacognition in LLMs

We introduce existing metrics and experimental procedures to measure and elicit metacognition in LLMs.

Psychologically-Grounded Measures. In cognitive psychology, confidence is considered an overt indication of behavioral uncertainty whose alignment with task performance is well-established as a measurable indicator of metacognition [85, 87, 84, 83]. A prominent paradigm for quantifying human metacognition is based on signal detection theory (SDT) [195, 13, 86], and it is designed to isolate metacognitive ("type 2") sensitivity (the efficacy with which confidence ratings distinguish correct and incorrect responses) from confounding factors of response bias (tendency to be confident) and task performance ("type 1" sensitivity). Under this framework, meta-d' and d' are measures of metacognitive sensitivity and cognitive ability, respectively, and their ratio $\frac{\text{meta-}d'}{d'}$ (M-ratio) and difference meta-$d' - d'$ (M-diff) represent metacognitive efficiency by normalizing metacognitive sensitivity relative to type-1 task sensitivity.[2] Recent work has also proposed information-theoretic alternatives to SDT, such as meta-$\mathcal{I}$ and relative metainformation, which quantify how much information confidence ratings carry about response accuracy while reducing reliance on parametric SDT assumptions [61, 205].

Several works have proposed methodologies inspired by SDT to quantify the metacognitive ability of LLMs. Such approaches generally pair meta-d' with a specific procedure to elicit model responses and associated confidence scores [243]. For example, Trinh et al. [278] use meta-d' to quantify how reliably a model's confidence (estimated as the maximum softmax output probability) predicts its own accuracy to inform test-time model selection, and Park et al. [220] utilize a dual-questioning protocol to elicit answers and self-evaluations of knowledge from LLMs to estimate meta-d'. Wang et al. [288] adapt meta-d'/d' logic to LLMs with Decoupling Metacognition from Cognition (DMC), in their experiments converting multiple-choice items into binary-choice tasks and eliciting answer confidence. DMC estimates task sensitivity $\hat{d}$ and computes $MC = d^*/\hat{d}$, making it closely analogous to the M-ratio while aiming to separate confidence-based failure prediction from raw task performance. Other psychology-based methodologies include the use of judgments of learning (i.e., self-predictions of future memory performance) [117] or vectorized dimensional measurements of metacognitive capabilities of LLMs [244]. While these approaches represent meaningful progress in measuring

[1] Nonetheless, improved capacity for metacognitive monitoring and regulation may enable systems to strategically modify outputs or internal signals to evade oversight or pursue unintended objectives (e.g., Li et al. [162]); we discuss such risks further in §8.

[2] An M-ratio of 1 indicates that an individual is an optimal metacognitive observer whose confidence captures all the information available from type 1 evidence. On the other hand, $M < 1$ indicates metacognitive loss, wherein the confidence signal is less informative than expected based on the evidence, and $M > 1$ indicates the confidence signal assesses information beyond what is driving the observer's type 1 judgments.

LLM metacognition, they are not without their limitations, including dependence on task-specific design or constrained task formats, sensitivity to confidence elicitation method, and lack of empirical justification. Moreover, SDT-based approaches typically require constrained response formats or external correctness judgments, making their extension to open-ended generation less direct.

Neurofeedback-Based Measures. Neurofeedback is an established neuroscientific technique in which participants are tasked with regulating their brain function in response to real-time biological feedback signals [255]. A classical example is in fear reduction experiments: subjects are presented with fear-inducing images, provided with a fear score derived from real-time recordings of their brain activity, and asked to modulate their neural activity to reduce the neurofeedback score. In a similar fashion, select works have sought to simulate a neurofeedback environment for LLMs to study their ability to monitor and control internal activations. This presents an alternative metacognitive evaluation route that does not depend on behavioral outputs (e.g., self-reported confidence scores may misrepresent actual uncertainty of a model). For example, Li et al. [162] devise a neurofeedback approach based on in-context learning wherein models participate in a multi-turn dialogue and are tasked with either implicitly or explicitly modulating[3] internal activations, while Yalon et al. [307] explore a similar setup to test models' ability to metacognitively predict the dominance of internal beliefs. Similar to psychologically-inspired methods, neurofeedback-based methods must also disentangle metacognitive from cognitive processing.

Confidence-Based Measures. Another class of approach directly uses the quality of judgments of intrinsic confidence as a proxy for LLM metacognition. Such works mainly focus on large reasoning models: Ma et al. [192] inspect the alignment between step-wise intrinsic confidence and process-level quality labels, quantified via typical calibration measures such as area under the receiver operating characteristic curve (AUROC) [24] and area under the precision-recall curve (AUPR) [196], while Lu et al. [189] define metacognitive confidence to estimate a model's belief about its own knowledge, which may be miscalibrated with factuality and help to explain high-confidence hallucinations. These methods are only loosely based in psychology and do not discuss potential bias relative to type 1 task performance.

Interpretability-Based Measures. As metacognition concerns the ability to introspect on internal states, other works have turned to interpretability techniques to detect signatures of metacognition in LLMs. This includes use of task-specific metacognitive probes to investigate whether correct and incorrect responses or reasoning traces are associated with statistically distinguishable internal computational signals (e.g., last token representations, self-attention scores, fully-connected activations, final layer softmax probabilities) [165, 160] and measurement of models' ability to accurately detect the presence of injected concepts [177]. While such analysis depends on access to model weights, it is generally extensible across models and task settings.

Task-Specific Measures. Metacognition is a fundamental capability that operates within and across domains, but in the early pursuit of eliciting and understanding metacognition in LLMs, it can be advantageous to focus instead on targeted, task-specific evaluations. To this end, some studies have designed novel experimental procedures to examine metacognitive capabilities of models in controlled settings. For example, Ackerman [1] introduce two experimental paradigms inspired by research on metacognition in nonhuman animals, wherein the ability for models to strategically deploy knowledge of internal states is tested without relying on self-reports; this includes assessment of models' ability to identify their knowledge (i.e., do models know *what* they know) and of models' certainty of their knowledge (i.e., do models know *that* they know), in a game-like setup. To evaluate LLM metacognition in dynamic contexts more reflective of real-world applications, Prasad and Nguyen [224] utilize a zero-sum multi-turn debate setup in which models are required to update beliefs and perform accurate self-assessment in extended interactions. Other efforts focus on investigation of specific metacognition-inspired capabilities, including selection and organization of task-specific skills [62], self-detection of biases during essay writing [280], reflection on creative generation tasks [55], self-prediction of generation temperature, and metalinguistic analysis [15]. In general, such approaches rely on carefully crafted prompts to elicit and inspect models' higher-order metacognitive abilities.

[3]The ability for models to report on neural activations is also explored.

Benchmarks for LLM Metacognition. Beyond individual experimental procedures, it is possible to construct standardized benchmarks to more holistically evaluate the metacognitive competence of LLMs. Several recent benchmarks measure metacognition in LLMs by evaluating their ability to reason about their own knowledge, uncertainty, or intentions [28, 211]. CogEdit [76] evaluates metacognitive behavior during knowledge editing by constructing questions involving counterfactual, boundary-constrained (appropriate generalization without overgeneralizing), and noisy editing (filtering potentially unreliable information) scenarios, to capture self-awareness and reflective reasoning during model updates. MetaMedQA [100] tests whether LLMs can recognize knowledge limitations and detect unanswerable questions in medical QA settings, by tasking models with making such metacognitive judgments on multiple-choice questions containing fictional concepts, malformed questions, and altered answers. ObjexMT [135] measures metacognitive calibration by asking models to infer the latent objective of multi-turn conversations and report confidence in that answer, with calibration measured via ECE and high-confidence error rates. Finally, AwareXtend [221] evaluates self-awareness and social awareness of LLMs across five dimensions of capability, mission, emotion, culture, and perspective, by observing performance on multiple-choice questions testing these abilities, and having several LLMs collaborate via voting or debate to produce the final answer. Together, these illustrate diverse aspects of metacognitive behavior that can be benchmarked in LLMs. Nevertheless, they remain constrained to specific task settings, disjoint in the types of metacognitive skills assessed, and subject to influence of general ability.

Gaps in Metacognitive Evaluation. To date, higher order skills such as metacognition and creativity remain under-represented among LLM benchmarks, which tend to focus instead on evaluating lower-level cognitive capacities [115, 321]. This is partly driven by the difficulty of reliably and consistently measuring metacognitive performance, evidenced by the earlier-named limitations of the methodologies introduced above. Moreover, some analysis [27] shows that single-point confidence probes and post-hoc calibration techniques typically used in metacognitive assessment of LLMs exhibit consistent failure modes. This can potentially be explained by the impact of post-training procedures such as reinforcement learning with human feedback (RLHF), which can systematically alter models' internal uncertainty (e.g., change the underlying probability distributions) and thereby influence the connection to models' output behavior [331]. Despite this, the incorporation of metacognitive evaluations into future benchmarks is important, as it can enable more accurate assessment and enhancement of model capabilities and inform oversight and safety. Especially as LLMs assume greater agency in real-world applications, the observation of consistent, stable, reproducible metacognitive phenomena in LLMs in certain settings (§4.2) warrants further investigation and inquiry. This motivates the development of more robust and systematic methods to measure, elicit, and evaluate metacognitive capabilities of LLMs across diverse contexts.

It is worth noting that while LLMs may exhibit behaviors that suggest a form of "artificial metacognition" (e.g., assigning calibrated confidence scores to outputs, adapting outputs based on reflective feedback, performing self-referential inquiry), the autoregressive nature of such models has led to the view that they present no real (meta-)cognition to speak of [115], with apparent metacognitive tendencies simply arising as a byproduct of LLMs' replication of complex behavioral patterns derived from their training data [266]. To this end, some have argued for more formal comparisons between human and LLM metacognitive processes [222]; this can help to clarify whether models are simply simulating superficial aspects of natural metacognition, and inform the development of more human-like and effective intelligent systems.

4.2 Current Findings on Metacognition in LLMs

While LLMs have shown great proficiency in various cognitive tasks, their ability to engage in metacognitive processes remains underexplored. In this section, we summarize the key findings and implications of current studies on metacognition in LLMs. We organize our discussion according to the type of metacognitive ability explored, roughly sorted from general to specific, and follow it by discussing the impact of factors such as model size, post-training, and confidence estimation method.

Metacognitive Sensitivity and Calibration. Given the nascent state of research on LLM metacognition, a number of studies have aimed to simply profile the baseline metacognitive performance of LLMs. Beyond the findings of measurement-focused works such as those discussed in §4.1, which establish that metacognitive sensitivity and efficiency of models are generally weak to mod-

erate, analysis of proprietary LLMs [222, 36] suggests that such models can, in fact, achieve better metacognitive sensitivity than humans on specific tasks (e.g., predicting outcomes of American football games), although such experiments tend to use small sample sizes (e.g., < 50 examples), limiting generalizability. Findings on metacognitive calibration are mixed and task-dependent: while Pavlovic et al. [222] show models tend to be underconfident on a coaching task, Cash et al. [36] find across several tasks that LLMs are largely overconfident—a phenomenon well-documented[4]—and struggle to retrospectively adjust confidence judgments based on past performance, demonstrating a metacognitive limitation not observed in humans. Metacognitive deficiencies of open-source models are likewise apparent, with Prasad and Nguyen [224] identifying five consistent metacognitive failures of leading open-source and proprietary LLMs in a selection of debate-focused and multi-turn task settings, including initial overconfidence, escalating certainty, mutually impossible high confidence, self-debate bias, and misaligned private reasoning. Sha et al. [245] further show that metacognitive sensitivity can be harmed by reasoning, with long reasoning traces hindering accurate self-assessment despite boosting performance, and distillation of large models' reasoning traces into smaller LMs impairing metacognitive sensitivity. Finally, Li et al. [172] examine the impact of declarative and procedural knowledge on model performance in a range of task settings, as these directly inform metacognitive processing in humans. They find that declarative information is generally more beneficial, while procedural knowledge is more useful in simple logical reasoning tasks, and that models' ability to leverage both types improves during pre-training at rates that vary with training step and model size. Overall, while LLMs can somewhat successfully demonstrate functional metacognition, substantial gaps remain, presenting potential risks in downstream agentic and decision-making applications.

Metamemory. The ability to introspect on memory is another core aspect of metacognition which has received attention [173]. Huff and Ulakçı [116], Huff and Ulakci [117] study the ability for GPT models to predict their own future memory performance, known as judgments of learning (JOL) in humans. They find that models largely fail at this task, signaling a lack of associated monitoring mechanisms, and underscore the need for further advancements in LLMs' self-monitoring capabilities, particularly to assist pedagogical applications. However, it is unclear whether such limitations extend to non-GPT models or more recent GPT variants.

Faithful Communication of Uncertainty and Decisions. Another promising direction of metacognitive study aims to understand the extent to which models can understand and faithfully explain their own internal processes. For example, the ability to detect different sources of uncertainty at play when responding to a query, as well as the decision-making processes involved, and accurately communicate these to a user is important and aided by metacognitive awareness. To this end, several works [316, 180] have turned to study the faithfulness with which LLMs express their intrinsic uncertainty—also known as *faithful calibration*. The aim is to align models' expressed and intrinsic uncertainty, where expressed uncertainty may be formulated numerically (e.g., "90% confident") or embedded into the model's output via use of linguistic uncertainty markers (e.g., "It is likely that..").[5] Yona et al. [316], Liu et al. [180], Liu and Cohan [179], Liu et al. [181] recently demonstrated that both open- and closed-source frontier models largely fail to faithfully express their intrinsic uncertainty, even with the application of specialized prompts or existing calibration approaches. Despite this, metacognitive prompting—wherein a system prompt is used to instruct models that they possess good metacognition—and the use of *metacognitive accuracy* as an *additional* training signal during RL were found to consistently and significantly improve faithful calibration of diverse LLMs, and the quality and helpfulness of models' resulting uncertainty expressions. This suggests that enabling models to better assess their own task performance can be important in expanding model capabilities. On the other hand, Cash and Oppenheimer [35] study the ability for LLMs to explain their reasoning processes and the weight assigned to different factors in decision-making. They find that frontier LLMs such as GPT-5 can, in fact, hold and express metacognitive knowledge of their decision processes to a degree comparable to humans, but that this is a relatively new capability not observed in earlier versions such as GPT-4o. As faithful explanations measurably influence user

[4] Systematic overconfidence is generally observed in a task-agnostic fashion in smaller models, and in difficult task settings in larger models. It can be attributed to a number of factors, such as RLHF post-training [155], task length [187], and use of imbalanced datasets in which failure cases or uncertainty rarely occur [331, 51, 259] in the calibration literature [295, 304, 49]

[5] Faithful calibration thus differs from the traditional view of calibration which instead seeks to align confidence with accuracy; the two calibration dimensions may diverge, e.g., when a model's internal uncertainty deviates from its accuracy.

trust in AI applications, enabling models to transparently disclose internal processes is central to appropriately calibrating user reliance on LLM outputs [262].

Self-Prediction of Task Performance and Difficulty. Beyond contributing to transparency, metacognitive monitoring enables humans to self-assess performance and predict task difficulty, leading to more efficient and effective problem-solving via subsequent metacognitive regulation. Existing research suggests that while LLMs can engage in similar monitoring behaviors, the efficacy of such judgments is limited and generally does not translate to reliable strategic adaptation [29]. For example, Barkan et al. [12] demonstrate that while leading LLMs can predict task success with better-than-random discriminatory power, such predictions do not improve with task familiarity, nor do reasoning models outperform non-reasoning LMs in making such judgments. Hwang et al. [118] show that reasoning ability does not correlate with such metacognitive awareness, even when estimating the complexity of simple reading comprehension tasks. Analysis by Binder et al. [19], Li et al. [160] further reveals that prediction of task difficulty is particularly challenging for models in hard or out-of-distribution settings,[6] and that even when predictive metacognitive information is present, it often fails to lead to effective monitoring and behavior adaptation. Lastly, overconfidence was found to lead to poor decision-making in agentic settings, providing evidence for the negative impact of models' poor awareness of their capabilities on task performance. This motivates the acquisition of a deeper understanding of models' self-assessment capabilities. Yet it is also important to consider potential risks of improved self-monitoring, including potential subversion of oversight mechanisms (§8).

Resource Allocation and Tool Use. Metacognition supports efficient allocation of computational and informational resources during learning and task execution. Yet LLMs lack the ability to reason about how to best allocate limited effort across multiple problems or options, including limited ability to estimate potential returns on allocated effort or to decide when to persist (e.g., continue searching) or move on [327]. While Li et al. [165] show that a probe trained using metacognitive information can improve tool use decisions, this relies on external interventions rather than judgments by a model itself. Work in this sub-area is limited, and further investigation can be worthwhile, especially toward the development of more efficient agentic systems and effective tool use.

Awareness of Knowledge Limitations. As it is generally desirable for a system to be able to recognize the scope of its own knowledge—to avoid misalignment between perceived and actual capabilities, improve reliability, and enable adaptive tool use or abstention—another line of research has sought to profile the degree to which LLMs embody this metacognitive attribute [225]. Findings in this area are generally mixed, with some reporting a lack of such ability in LLMs [315, 100, 129, 291], and others suggesting models can exhibit functional access to knowledge of internal states [225, 19, 55], but often to a lesser degree than humans [315]. Ackerman [1] show in a related vein that models can sometimes anticipate what answers they will give and use that information to demonstrate self-knowledge. Park et al. [220] also demonstrate that LLMs can be trained to introspect on their own knowledge and reference internal knowledge during generation rather than exhibiting superficial alignment. Such insights form an interesting basis for further study, which may look toward more systematically characterizing such abilities of LLMs. For example, it is of interest to understand the extent to which LLM-based systems can leverage their recognition of insufficient or outdated internal knowledge to improve at related abilities such as abstention or the selection and evaluation of different tools [146]. Such behavior mirrors the phenomenon of *metacognitive control* in psychology.

Introspection: Monitoring and Reporting Internal States. Separate from recognition of self-knowledge, a parallel strand of work has inspected models' ability to introspect on internal states. Introspection is defined as the process of acquiring or accessing knowledge that originates from internal states and is not accessible to external observers [19]. Work in this direction is likewise sparse, but has taken on a few distinct directions. First, Li et al. [162], Yalon et al. [307] study the ability for models to monitor and report their own belief states based on exemplars, finding that both large (up to 70B parameters) and small models ($\sim$8B parameters) often perform above chance at such tasks and can even control activation patterns in some settings. Yalon et al. [307] present causal evidence

[6] Despite this, in certain settings it appears that models can effectively assess and utilize their own confidence in answering factual and reasoning questions—for example in a synthetic game setup designed by Ackerman [1]. This provides additional evidence suggesting such self-assessment ability may be task-specific.

for the ability, and Li et al. [162] further show that the space spanned by the directions of neural activations able to be reported or controlled in such a fashion is of significantly lower dimensionality than a model's neural space, suggesting ability to monitor only a subset of activations covered by this "metacognitive space." Still, this capability appears to be partially mediated by the number of in-context exemplars used for the monitoring and reporting tasks, the semantic interpretability of the studied neural directions, and the amount of variance explained by such directions. A second direction of study examines whether models can identify the presence of concepts injected into activations [103, 177]; it is shown that models can do so in certain scenarios, indicating a degree of functional introspective awareness of internal states. However, this ability is fragile. A final avenue of study specific to knowledge of language is presented by Song et al. [256], who, in contrast, provide negative conclusions regarding models' introspective ability. We refer the reader to that paper for further discussion of potential explanations. More generally, the ability for models to introspect on internal states is highly context-dependent, with the factors governing it not yet fully elucidated, and presents important implications for the robustness of oversight mechanisms.

Behavioral Self-Awareness. Models' capacity to accurately describe learned behaviors without explicit examples or supervision is another important metacognition-derived capability [17]. Limited work suggests models trained to exhibit specific behaviors can articulate signatures of self-awareness across diverse tasks and contexts, and that such awareness is acquired early on in training and can be mechanistically localized to a single steering vector in a model's activation space [25]. Other related investigation includes examination of introspective awareness, defined as knowledge of identity, capabilities, and objectives [221], which can be improved by collaborative reasoning among multiple agents.

Reasoning. The ability for reasoning models to engage in reflective and metacognition-like behaviors [65, 91, 66, 90, 136, 303] has prompted questions regarding whether the acquisition of reasoning capabilities is accompanied by associated gains in metacognitive performance; this has led to an active line of inquiry into the connections between metacognition and reasoning in LLMs.[7] To this end, several studies have provided evidence for limited metacognitive capabilities in LRMs, finding that such models struggle to leverage metacognitive information to engage in monitor or control behaviors [160] and that enhanced reasoning capability can act in opposition to metacognitive sensitivity—demonstrating improved metacognition is not an automatic byproduct of better reasoning [245]. Marjanović et al. [197] show that DeepSeek-R1 is incapable of simple metacognitive monitoring tasks such as detecting the length of its own reasoning traces. Despite this, distilled variants such as DeepSeek-R1-Distill-Qwen32B and DeepSeek-R1-Distill-LLaMA-8B have been observed to maintain reliable internal estimates of relative position in reasoning, suggesting a fundamental difference in metacognitive behavior and potentially also in the acquisition of such abilities [71]. Lastly, the impact of metacognition on reasoning efficacy is explored to some degree by works such as Lu et al. [189]; such study reveals that chain of thought faithfulness to the final answer often diverges in settings where models exhibit weak metacognitive sensitivity. It is also observed that the impact of reflective behaviors on model confidence is often minimal, suggesting a lack of metacognitive grounding. These findings underscore the need to move beyond surface-level alignment in LRMs. Moreover, metacognition remains a relatively understudied component of LRM research, with only a small portion of related papers considering metacognitive aspects [130].

Task-Specific Findings. Finally, a range of recent studies have yielded insights regarding the metacognitive skills of models in task-specific functional contexts. This includes assessment of models' ability to: name and strategically apply skills to math reasoning tasks via guided prompts [62], deploy metacognitive strategies to perform knowledge editing [76], self-reflect during essay writing [280], introspect on generation temperature [55, 257], perform metalinguistic analysis [15], self-assess and revise positions in adversarial debate [135, 224], and execute system prompt learning [219]. Findings regarding these metacognition-derived capabilities are generally mixed across models and task contexts. Nevertheless, these settings constitute a rich testbed for probing and evaluating LLMs' metacognitive abilities in a manner more representative of the diversity of human metacognitive faculties.

[7] For example, Chen et al. [47] find that improving a model's self-verification ability leads to improved reasoning performance.

Impact of Model Size, Family, and Performance. Existing works generally uncover a consistent association between model size and metacognitive ability [288, 177, 220], regardless of whether the focus of study is metacognitive efficiency, metacognitive sensitivity, or introspective awareness. This suggests metacognitive behavior may naturally emerge alongside the enhanced performance and generalization of larger models. Despite this, such findings are supported by only a limited collection of datapoints, obtained from either small, sub-9B models or large proprietary models of undisclosed size, and from specific, disjoint task settings across model scales; it is unclear whether the observed metacognitive patterns persist in moderately-sized LMs and in broader task settings and model families, including proprietary LLMs. Additionally, as noted by Lindsey [177], such trends can be complex and further sensitive to other factors such as post-training strategy, which we discuss below.

Model family presents another dimension of variation affecting metacognitive analysis of LLMs. Cacioli [26] find that metacognitive efficiency varies substantially across models even when task performance is similar, that models with similar calibration level (measured via expected calibration error) likewise exhibit distinct gradations in metacognitive efficiency, and that the impact of temperature on metacognitive sensitivity and efficiency can vary markedly by model family (e.g., Mistral and Gemma 2 exhibit reverse trends according to generation temperature). Park et al. [220] further demonstrate that among models of similar sizes but from different families, clear rankings according to metacognitive performance can be derived.

The use of reasoning models can introduce additional complexity that causes trends to diverge from those observed with non-reasoning LMs. In particular, LRMs with better overall performance do not necessarily exhibit the strongest metacognitive sensitivity [245], and in some model families the smallest reasoning models exhibit higher metacognitive sensitivity than their larger counterparts. Sha et al. [245] attribute this to the profuse use of intermediate reasoning steps in larger models, which helps performance but increases the difficulty of accurate self-evaluation; however, their study focuses exclusively on digit manipulation tasks. On the other hand, Advani [3] study the task of metacognitive reflection and suggest that there exists a "capacity threshold" of at least 7-9B for reasoning models under which metacognitive prompting interventions can be harmful. They find that while small models can identify and correct some reasoning errors via such reflection, they struggle to avoid introducing new errors during this process and often instead enter into hallucinations or logical leaps—concluding that metacognition leads to confusion without sufficient model size or capacity. It is worth noting, however, that this work defines metacognition from a conceptual standpoint that is less grounded in technical definitions seen in psychology. Huang et al. [111] present similar findings in a non-reasoning setting that smaller models are more resistant to metacognitive prompting; yet, this may be simply due to weaker instruction-following capacity.

More generally, the lack of consistency in evaluation approach, task domain, and experimental setup (prompt, confidence elicitation approach, etc.) makes it difficult to reliably compare and validate findings and trends regarding models' metacognitive faculties. This remains an open area for systematic, comprehensive investigation.

Impact of Post-Training. The impact of post-training procedure on metacognitive ability is not yet well-established. Lindsey [177] provide preliminary evidence, in addition to their commentary on the impact of model size and performance, that introspective strength is sensitive to post-training approach. Cacioli [26] also directly compare corresponding instruct and base models from the Llama 3 family, finding a domain-specific effect in terms of the impact of post-training on metacognitive efficiency. Namely, while both variants exhibited nearly identical meta-d' and d', the M-ratio was slightly decreased for the instruct model, an effect more pronounced in the scientific domain versus art, literature, history, politics, and geography. They conclude that RLHF may specifically degrade metacognitive efficiency on STEM tasks. Additional analysis reveals that across different temperatures, while meta-d' and d' of base models remains fairly stable, instruct models exhibited a dissociation in which one of the two quantities increased or decreased while the other was stable (pattern specific to model family). Such findings point to a complex dynamic between post-training and metacognitive ability which remains to be fully profiled, despite its importance.

Impact of Sampling Temperature. Generation temperature can have a measurable impact on resulting metacognitive observations. While the impact of temperature is well-recognized in the

calibration literature,[8] complementary findings for type 2 tasks are only just beginning, with Cacioli [26] showing that temperature manipulation leads to divergence between metacognitive sensitivity and confidence policy for instruction-tuned models. It is observed that if a model's metacognitive efficiency is too low, calibration via temperature adjustments cannot meaningfully address underlying metacognitive deficits, with confidence signals remaining insufficiently informative. On the other hand, base models which do not exhibit significant changes in confidence or sensitivity as a result of temperature setting do not seem to reflect this trend. The veracity and generalizability of such conclusions remain to be studied, warranting further investigation into the role of temperature in shaping demonstrated metacognitive performance of LLMs.

Impact of Confidence Estimation Method. A potentially significant source of inconsistency among existing metacognitive evaluations of LLMs is the methodology used to estimate model confidence [288]. For example, response-level confidence may be estimated via sampling consistency [194, 14, 41, 131, 304], semantic variability [203, 145, 99, 214], token probabilities [126, 67, 114], or self-reported confidence scores [122, 272, 108, 306, 309, 328], among other approaches. While systematic study of the impact of this decision is largely absent, we summarize a few relevant insights from recent work. First, Dai [60] study the impact of verbalized confidence scale manipulations on metacognitive sensitivity, finding across leading open-source and proprietary LLMs (e.g., GPT-5.2) in long and short-form tasks that confidence scale design directly impacts the quality of self-reported confidence scores and thereby also metacognitive performance. In particular, they observe that verbalized confidence values elicited in the typical 0–100 range are highly sparse and discretized, with over 75% of responses concentrating on three values, and that alternative use of a 0–20 scale markedly and consistently improves metacognitive efficiency. To bypass this issue, Wang et al. [288] leverage token log-probabilities to estimate model confidence, focusing on sub-9B open-source models in short-form QA tasks. They find that with use of such continuous confidence estimates, M-ratios approximately ranging from 0.85–1.05 are obtained, in contrast to M-ratios of approximately 0.62–0.92 obtained with the use of self-reported confidence scores [60]. Further evidence for the variability of metacognitive evaluation results with confidence method are provided by Wang et al. [288], whose proposed SDT analog for LLMs exhibits similar sensitivity.

What Metacognitive Metrics Capture that Calibration Metrics Don't. Theoretically and empirically, analysis of metacognitive performance offers additional insights regarding the quality of model confidence signals which may not be revealed through typical calibration-based evaluation alone. Consider the setting in which multiple models appear to be well-calibrated according to metrics such as ECE [101], Brier Score [97], or AUROC [260].[9] In this case, metacognitive sensitivity specifically distinguishes among models which "know what they don't know," which can be important for model selection in confidence-critical downstream settings and impact human-AI interactions. For example, if two models with the same ECE exhibit different d' and meta-d' behavior, and one model is to be deployed to a downstream system for selective prediction tasks, then the model with higher d' (task performance) but lower M-ratio (metacognitive efficiency) would underperform relative to the model with lower accuracy but higher metacognitive efficiency. This tradeoff is not reflected by ECE alone. Cacioli [26] cite such conceptual examples along with empirical evidence that AUROC and M-ratio yield fully inverted model rankings, to posit that confidence-dependent applications should therefore prioritize high M-ratio models. Metacognitive evaluation of LLMs can therefore provide a complementary lens with which to assess model quality. Yet extending such evaluation to open-answer tasks—a key benchmark setting for LLMs—remains unexplored.

Domain-Generality vs. Domain-Specificity of LLM Metacognition. Whether human metacognition relies on a domain-general resource shared across different tasks or domain-specific evaluative

[8] That is, different temperatures can lead to different confidence estimates, especially if using sampling-based confidence estimation [194, 41].

[9] Such confidence metrics can be categorized into three tiers depending on their relevance for metacognitive assessment [26]: "tier 1" measures such as ECE and Brier Score are known to conflate sensitivity (how well correct and incorrect responses are separated) with bias (overall confidence level) and vary with discretization scheme, and "tier 2" measures such as AUROC and phi correlation are known to confound type 1 performance despite capturing ranking quality. Within these, the Brier score improves on ECE by decomposing into reliability, resolution, and uncertainty, but it does not separate the model's discriminative capacity from its metacognitive sensitivity when controlling for that capacity. Likewise, the AUROC improves on ECE by measuring ranking quality, but it conflates task performance with the quality of metacognitive monitoring over that performance. On the other hand, "tier 3" measures include meta-d' and M-ratio, which control for type 1 sensitivity and thereby isolate metacognitive sensitivity/efficiency.

processes has long been contested [207]. Parallel questions of interest may be formulated for LLMs. That is, are metacognitive behaviors of LLMs mediated by a single "global" mechanism or governed by individual, domain-specific processes? Is the metacognitive performance of LLMs—measured via metacognitive efficiency, for example—consistent and generalized across domains? Does domain-specificity depend on the particular metacognitive skill being assessed? To the best of our knowledge, no work has yet considered the plausibility of the first or third questions. However, Cacioli [26] supply early evidence in support of the second question that metacognitive efficiency of LLMs is, in fact, domain-specific, with different models exhibiting varied performance across task domains.

Overall, the findings highlighted in this section emphasize the importance of systematically evaluating the meta-cognitive abilities of LLMs in diverse experimental and task settings. While LLMs appear to exhibit fundamental metacognitive limitations, complicated by factors such as post-training and confidence estimation approach, substantial opportunities remain for further investigation, particularly in understanding how these abilities can be measured, elicited, and improved, with important implications such as those discussed in §3.

5 Giving LLMs Metacognitive Abilities

Building on prior observations of metacognition in LLMs, we now examine methods for implementing and improving metacognitive abilities of LLMs, LRMs, and agentic systems. These approaches span architectural and modular designs, prompting-based techniques, and training strategies (e.g., use of self-generated signals during RL [138, 314]), among others. Notably, the prospect of encoding metacognition into AI systems predates the emergence of LLMs, with early methods leveraging metacognitive principles to improve robustness of intelligent systems [238], synthesize general-purpose "meta-models" [32], and translate human metacognitive abilities into AI analogs [246], in addition to other goals [300]. We refer the reader to such works for details on broader efforts toward artificial metacognition.

5.1 Implementations of Metacognition in LLMs

Current efforts to instantiate metacognitive functions in LLMs generally utilize specialized frameworks or architectures.

Frameworks. Some studies have aimed to draw directly from metacognitive frameworks presented in cognitive psychology to develop analogous mechanisms for metacognition in LLMs. For example, Yan et al. [308] draw inspiration from dual-process theory [127] to propose a framework for metacognition-assisted reasoning in LLMs that involves self-awareness, monitoring, evaluation, and regulation. Pangu Embedded [39] is a similarly inspired framework that adapts from dual process theory to implement system 1 (fast, intuitive) and system 2 (slow, deliberative) processing for reasoning-optimized LLMs, combining use of distillation fine-tuning, reinforcement learning, and metacognitive prompts to confer models with more nuanced, efficient, and versatile reasoning. Oh and Gobet [217] propose the Monitor-Generate-Verify (MGV) framework, which serves as a computational translation of metacognitive theories by Flavell [81] and Nelson [212] and incorporates explicit monitoring to capture metacognitive experiences prior to generation as well as retrospective monitoring and verification. This approach is concretely implemented by Oh [216], who demonstrate its benefits toward effective reasoning. More generally, Scholten et al. [239] present a theoretical framework to explain the presence of five symptoms of metacognitive deficiency in LLMs as a byproduct of limited monitoring and control capabilities. Conway-Smith and West [57] further posit that equipping LLMs with human-like metacognitive abilities requires implementation of human-like procedural memory.

Other frameworks directly design methods to implement metacognitive functions for LLMs in specific task settings. For instance, another framework targeted toward reasoning tasks is Meta-R1 [64], which introduces explicit metacognitive capabilities into reasoning-equipped LLMs and demonstrates consistent improvements in performance and efficiency across models and tasks. To improve mathematical reasoning, Huang et al. [113] present a modular framework named MetaMath-LLaMA which combines a transformer-based metacognitive subtask scheduler with symbolic parsing and symbolic-neural computation. Regarding awareness of knowledge boundaries, Park et al. [220] propose Evolution Strategy for Metacognitive Alignment (ESMA), which is a framework to strengthen

the alignment between models' internal knowledge and explicit behavior. A framework to simulate metacognitive learning with self-awareness, boundary monitoring, and reflective thinking to step toward human-like knowledge editing is introduced by Fan et al. [76]. Lastly, Li et al. [169] propose a method for metacognitive test-time reasoning to enable vision LMs (VLMs) to self-reflect and learn adaptively during inference.

Architectures. To implement metacognitive functions in LMs, a few works have also proposed specialized model architectures. Aviss [8] introduce the State Stream Transformer (SST) architecture, finding that it can exhibit enhanced reasoning capabilities and emergent metacognitive behaviors, which they attribute to higher-order processes of planning and regulation. On the other hand, Jha et al. [121] present SAGE-nano, a 4B parameter model which extends the transformer architecture with specialized metacognitive components that enable inverse analysis of forward reasoning. Such architectures generally maintain or enhance reasoning performance and can improve model transparency. A related innovation in optimization is the Metacognitive Introspective Reward Architecture (MIRA) proposed by Zhang et al. [324], which utilizes a two-tier optimization loop to drive policy learning while incorporating metacognitive supervision to monitor learning dynamics.

More generally, much of the effort in this space focuses on implementing metacognitive functions to assist reasoning-centric applications of LLMs; while important, empirical support for these methods can be limited, and it can be valuable to devise similar methods for broader use cases.

5.2 Metacognition for Reasoning Models

A growing body of research has examined how metacognitive mechanisms can be incorporated into LRMs. Ha et al. [102], Wang et al. [289] find that LRMs already exhibit latent reflective behaviors commonly associated with expert metacognitive reasoning, such as self-critique and reflection, motivating further work to regulate and amplify these tendencies. Inspired by this, several prompting frameworks have been developed to explicitly elicit structured metacognitive reasoning. Shakoo and Sameti [247] introduce the Dynamic Cognitive Orchestrator (DCO), a two-stage framework that first generates a problem-dependent reasoning strategy and then executes it, yielding strong performance on complex reasoning tasks. Similarly, Fang and Chen [77] propose the Probabilistic Chain-of-Evidence (PCE) framework, which integrates uncertainty quantification into the reasoning process via prompt engineering, consistently outperforming classical CoT on metrics such as hallucination rate. Li et al. [160] adopt a complementary emergent approach, dynamically adjusting thinking strategy through separate planning and solving models to improve mathematical reasoning. A parallel line of work focuses on metacognitive reuse at inference time: Yang et al. [310] first introduced Buffer of Thoughts (BoT), which draws on distilled "thought-templates" from prior problem-solving processes to improve performance and generalizability. Didolkar et al. [63] extend this idea to reasoning models, integrating a "behavior handbook"—a record of behaviors curated from previous traces—that reduces token usage while maintaining accuracy. Cheng et al. [48] apply a similar principle within MCTS, using a shared memory of heuristics and fallacies drawn from metacognitive reflection over prior search trajectories to guide subsequent reasoning branches, halving trajectory requirements while improving accuracy. Xiang et al. [302] also target inference-time efficiency, monitoring reflection signals during generation to evaluate whether the current CoT is sufficient to terminate, reducing reasoning length by almost 35% with minimal accuracy loss.

On the training side, researchers have developed methods to embed metacognitive capacities within reasoning model parameters, broadly falling into two categories: fine-tuning on augmented traces, and integrating self-generated signals. Ha et al. [102] address the problem of "overthinking"—where models generate repetitive reasoning—by decoupling reasoning and control into independently optimized components, improving accuracy and reducing token usage. Wang et al. [289] apply SFT and RL on self-critiqued reasoning traces to improve both accuracy and reflection quality; Sun et al. [264] similarly fine-tune on self-corrected traces, adding a hierarchical decomposition step for harder problems. Li et al. [160] extend their prompting-based approach through SFT and RL on traces augmented with difficulty assessments and planning decomposition. Didolkar et al. [63] similarly incorporate their "behavior handbook" into an SFT objective, achieving improved token efficiency and accuracy. The second category integrates self-generated signals into the training objective itself. Kim et al. [138] fine-tune LRMs by aligning predicted statistics about forthcoming generations—such as pass rate and solution length—with actual outputs, significantly improving accuracy and out-of-domain generalization. Yeom et al. [314] instead incorporate self-signals

regarding answer correctness, improving both accuracy and calibration—a property of particular importance for LRMs deployed in mission-critical settings. A third form of approach bypasses reasoning trace supervision entirely: Leong et al. [156] show that LRMs possess latent reasoning beliefs that track their own reasoning traits, and leverage this by fine-tuning on metacognitive QA pairs that reshape the model's self-concept toward desired behaviors—matching stronger baselines at lower cost. Zhao et al. [327] takes a complementary angle, training models to estimate the expected value of additional reasoning before generation and enabling principled decisions about whether to invest compute in a given problem. Li et al. [164] propose mid-training on pairwise comparisons between reasoning traces to improve the reasoning-metacognition tradeoff. Finally, Li et al. [159] take an introspective approach with CoRE-Eval, analyzing the geometry of hidden-state trajectories during reasoning to identify cyclical patterns of redundant deliberation, enabling training-free early termination that reduces reasoning length and improves accuracy.

5.3 Metacognition for LLM Agents

A third line of inquiry concerns the implementation of metacognitive capabilities into agentic systems to increase their practical utility and tackle their limitations [323]. In particular, while LLM agents have gained traction in a variety of applications, they continue to exhibit deficiencies such as rigid task chaining, shallow reasoning, brittle memory retention, becoming trapped in unforeseen loops, or encountering unrecoverable failures, which can lead to user frustration and reduced trust [305, 5]. In this subsection, we discuss the works which have pursued research in this direction, categorizing them according to the specific task setting or metacognitive capability studied.

Learning, Introspection, Adaptation, and Self-Improvement. Modern agentic systems encompass a wide range of structures, including use of hybrid architectures combining dynamic goal formulation and memory with natural language reasoning and environmental feedback [37, 312, 251]. The incorporation of metacognitive monitoring and control mechanisms presents opportunities to further improve their functionality. To this end, a number of works have developed metacognitive frameworks for self-improvement, self-prediction of competence and learning progress, improved strategy selection, and monitoring of "thought" processes and actions by LLM agents [277, 95, 305, 109, 281]. Such works generally demonstrate that augmenting agents with metacognitive capabilities enables improved task success [305], more effective learning in open-ended goal spaces [95], more systematic reflection and use of procedural knowledge [109], more iterative and adaptive strategy selection in diverse contexts [277], better ability to handle novel scenarios [281], strengthened subtask decomposition skills [290], and automatic generation of agentic workflows [332].

Multi-Agent Systems. Similar to single-agent settings, metacognitive methods can improve the capabilities and utility of systems in multi-agent contexts. Efforts in this direction include the development of training-based or modular instantiations of "meta-thinking" [18, 286] and the implementation of fine-grained error detection and self-correction mechanisms [249] for multi-agent systems. It is observed that such approaches can address known limitations of multi-agent systems, including error propagation, limited flexibility and reliability in complex or high-stakes contexts, and shortcomings in reasoning, LLM-as-a-Judge, and mathematical tasks, among others.

Reasoning. Reasoning in LLM agents is another use case which has received attention. While most work to enable metacognition-assisted reasoning in LMs focuses on reasoning models directly (§5.2), some studies have devised agent-specific methodologies, including AutoCrit [236], which integrates agents for critique, reasoning, and monitoring in a feedback loop to mitigate inconsistencies and error propagation; RefRea [193], which utilizes a metacognitive module to compare reference and reasoning model actions and guide outcomes via reflection and environment-responsive adaptation; and Light et al. [175] which uses scaffolded in-context meta-reasoning examples.

Collaboration. Additional work has investigated the contribution of metacognitive mechanisms toward improved agent collaboration. For example, Zhang et al. [325] draw upon psychological theories of metacognition and theory of mind to emulate human-like social reasoning in multi-agent settings, while Yang et al. [311] present a paradigm for human-agent collaboration that leverages judgments by a metacognitive policy to determine when to seek assistance from a human expert.

Memory. As LLM agents rely on accumulated memory to tackle long-horizon and decision-making tasks, memory management is another metacognition-informed capability which presents opportunities for improvement. In this direction, Liang et al. [174] highlight that many agentic approaches rely on fixed memory representations at a single level of abstraction, which can hinder generalization. Thus, they propose to treat memory abstraction as a learnable skill, and introduce a metacognition-inspired approach which trains a separate model via DPO to learn to distill knowledge into representations with suitable format and at an appropriate abstraction level for reuse.

Search, Retrieval, and Tool Use. A final application setting which has been explored is the use of metacognitive monitoring or reflection to better allocate resources across search, retrieval, reasoning, and other tools in agentic applications. For example, Sun et al. [265] introduce a monitoring framework for deep-search agents to separate execution-level control from task-level reasoning, improving the robustness and practicality of such agents. Similarly, Chen et al. [44] present a metacognition-inspired framework to enable agents to dynamically assess when to reason with existing knowledge or seek new evidence, which helps to eliminate redundant retrieval. Musumeci et al. [209] also leverage metacognition-derived self-reflection mechanisms to enable agents to approach knowledge-based QA tasks in a more structured fashion and dynamically orchestrate query answering. Another related direction concerns temporal awareness, raising questions such as how an LLM might know that its knowledge is outdated to verify more recent (external) sources.

6 Metacognitive Methods to Improve Capabilities of LLMs

So far, our discussion has focused on direct observation and improvement of metacognitive capabilities of LLMs. A direction which is similarly worthwhile, especially toward realizing diverse downstream benefits of metacognition-like processing (§3), is the devision of functionally metacognition-grounded methods to improve and extend the capabilities of LLMs. Efforts in this direction can be broadly categorized by whether they aim to improve performance in specific task settings or simply improve the efficacy and efficiency with which learning occurs.

6.1 Improving Task Performance

We discuss the various metacognition-inspired approaches aimed at boosting the performance of LLMs in diverse tasks.

General Performance. As metacognitive processing is broadly applicable across domains, one class of work consists of methods aimed at improving the general performance and problem-solving capabilities of LLMs. Typically, models undergo post-training (e.g., instruction tuning) to improve adherence to human preferences and instructions and acquire strong zero-shot capabilities. Along these lines, some work has shown that using instructions aligned with models' cognitive processing or inference-time metacognitive prompts inducing self-reflection and self-evaluation can drive up problem-solving and reasoning abilities across multiple benchmarks [231, 293, 9, 321]. Prompt designs capturing different types of metacognitive behavior contribute differently to resulting improvements [231, 180]. Additionally, the ability for models to detect and explain wrong predictions in QA settings can be improved via multi-stage fine-tuning in which the model acquires experience by encountering its earlier predictions; this lends toward self-validation in a metacognitive fashion [188]. Yet, fine-tuning may not be required to improve the self-correction capabilities of LLMs: Tan et al. [268] propose the use of sparse sub-networks to capture decision processes toward better introspective monitoring. Ample space remains to further explore how we can construct and leverage metacognitive methods to improve general capabilities of LLMs; recalling discussion of the domain generality vs. specificity of metacognition in §4.2, however, it may be more effective to focus on setting-specific instantiations.

Confidence Calibration. The ability to assess and communicate one's own uncertainty is a defining characteristic of metacognition. Yet while LLMs can maintain internal uncertainty signals, their expressed confidence is often misaligned with both accuracy and their internal uncertainty. To this end, some works have sought to devise metacognitive approaches to improve models' ability to signal their confidence to human users in a reliable and calibrated fashion. In terms of the factuality-based alignment between confidence and accuracy, Steyvers et al. [261] leverage supervised

fine-tuning (SFT) to strengthen models' ability to assign calibrated confidence scores to answers and discriminate between two possible answers based on anticipated correctness. They observe that multi-task training is necessary for effective generalization of this skill. On the other hand, another strain of work has investigated metacognitive methods to strengthen the faithful alignment between models' expressed and intrinsic confidence (i.e., "faithful calibration"). Liu et al. [180] show that metacognitive system prompts that directly tell a model it has strong metacognitive sensitivity can lead to consistent, moderate improvements in faithful calibration in a generalized fashion. Building upon this, Liu et al. [181] demonstrate that the use of metacognitive performance as an additional training signal during RL to scale advantage scores and preferentially refine completion rankings can lead to strong gains in faithful calibration while simultaneously conferring models with improved ability to self-predict performance in a point-wise fashion. This reinforcement learning with metacognitive feedback (RLMF) paradigm can outperform standard RL and generalizably improve performance at metacognitive monitoring and target tasks. This contrasts the SFT-based findings regarding generalizability by Steyvers et al. [261], suggesting the presence of more complex dynamics governing the generalization of metacognitive capabilities—a direction of study which is unexplored as of yet. Different types of metacognitive tasks (e.g., expressing confidence to comparing question difficulty) exist, and it is possible that some forms of generalization are easier to master than others.

Hallucination Reduction. The ability to self-identify errors and detect and express sources of uncertainty is characteristically well-suited for application toward hallucination mitigation. Preliminary work indeed suggests that endowing models with metacognitive capabilities—for example, by using average metacognitive error as the loss function for gradient updates during fine-tuning [166]—can reduce the occurrence of hallucinations during text generation. This method is conceptually similar to the RLMF paradigm proposed by Liu et al. [181], but uses a different measure of metacognitive performance, applies it at a different point in the training process, and focuses on fine-tuning without use of RL. It also involves a more complex design with multiple models. Nevertheless, this provides additional evidence for the value of leveraging metacognitive performance signals to inform training and optimization. Beyond training, Miki and Vincent [206] also demonstrate that metacognitive prompts to enforce structured system 1 and system 2 "thinking" in LLMs can reduce contextual hallucination, supporting the parallel approach of using prompts to guide models to exhibit metacognition-like behavior and subsequently improve output quality.

Knowledge Boundary Detection. The ability for a model to detect and signal the scope of its own knowledge is another important skill which can inform efficient resource use, improve factuality, and boost trustworthiness [225, 42, 191]. Yet LLMs often experience knowledge-confidence gaps which lead to overconfident errors or uncertain truths [40]. To address this, several studies have proposed the use of metacognitive methods to improve knowledge gap awareness. For example, Tiwari and Gupta [275] devise metacognition-inspired reflective prompts to elicit self-evaluation and error correction in this direction, which also has applications in hallucination mitigation. A more systematic approach is introduced by Chen et al. [40], wherein a model's internal signals are used to partition its knowledge space into regions of mastery, confusion, and absence and thereby inform targeted augmentation of existing knowledge. Roy and Roy [235] further leverage physics-based constraints to improve LLMs' self-knowledge ability and their performance at selective prediction, calibration, and response revision.

Resistance to Persuasion. In humans, a natural capability closely related to accurate confidence assessment and self-knowledge is the ability to resist persuasion when appropriate. Investigation of this property in LLMs as it relates to metacognition is limited, but preliminary evidence from Huang et al. [111] suggests that metacognitive prompting based on the elicitation of self-reported confidence scores can, in fact, increase vulnerability to belief erosion, and that model size, family, and post-training may be important in affecting this susceptibility. Despite this, it is important to note that verbalized confidence in LLMs is known to suffer from limitations of inconsistency and unreliability. We interpret this as motivation for further research into the potential benefits of metacognition toward robustness against undue persuasion or counterfactuals, wherein alternative metacognitive prompting, training, or other methodologies could yield different conclusions, and perhaps offer an alternate path toward improving resilience to misleading or adversarially crafted inputs.

Interpretability. A byproduct of improved ability to monitor and report on capabilities, knowledge, and confidence is enhanced interpretability. This view is posited by Tankelevitch et al. [270], who

suggest that existing explainability approaches can be augmented by considering metacognition. Empirical support for this proposal is somewhat limited, but works such as Plunkett et al. [223], while not explicitly concerned with metacognition, highlight that models' ability to report detailed, quantitative justifications of their decision processes can be improved through training, offering steps toward more generalized ability to describe internal operations in a way understandable to human users.

Reasoning for Non-Reasoning Models. Another concrete use of metacognitive methods for LLMs is to improve the reasoning performance of models not necessarily optimized for reasoning tasks. Works in this direction generally focus on a few specific classes of approaches. Some studies present novel frameworks to adapt principles of human metacognition for LLM reasoning. For example, Haque [104] present a metacognitive controller that enables models to dynamically select how to think: in effect, their framework uses reasoning about reasoning to identify the most suitable reasoning style (e.g., chain of thoughts, tree of thoughts, graph of thoughts) for a given input, shown to be effective for the FlanT5-Small model ($\sim$80M parameters). Other notable framework-based approaches include Huang et al. [113], Oh [216], Ta et al. [267], which focus on enhancing mathematical reasoning abilities of LLMs. These methods can improve the explainability of generated reasoning (e.g., clearer steps in solutions), beneficial toward aiding students and educators. Elenjical et al. [72] is another noteworthy work in this space which is more psychologically grounded. A second strain of work utilizes special metacognitive prompts to enhance reasoning performance. This includes instructing models to reflect on certainty of reasoning steps and step content [227, 77], identify knowledge conflicts before reasoning [10], and dynamically select among reasoning methods based on task requirements [93]. Such approaches often outperform chain-of-thought prompting and yield better uncertainty expression. A related technique leverages iterative reflection to induce metacognition-like processing during reasoning tasks [183], which can be particularly helpful for small LMs [163]. Khandelwal et al. [134] show that metacognitive monitoring of performance and subsequent provision of feedback can enable progressive solution refinement, leading LMs to outperform reasoning models. Finally, metacognition-inspired analysis can be used to model LLM reasoning via influence graphs [58], helping to identify critical points in reasoning traces and improve reasoning efficiency and reliability.

Retrieval. As resource allocation is a key function aided by metacognition, some have devised metacognitive methods to improve the efficiency of retrieval-augmented generation (RAG) procedures [165]. For example, Chen et al. [44] propose frameworks to dynamically evaluate exploration utility, identify coverage or relevance deficiencies, and assess when to seek new evidence or reason. This helps to eliminate redundant retrieval steps and improve context evolution and conciseness, leading to gains over text- and knowledge graph-based RAG baselines. Detection of knowledge boundaries is another metacognition-derived ability which Lv et al. [191] utilize via in-context learning to increase the accuracy of self-knowledge assessment prior to retrieval. General application of metacognitive functions such as monitoring, evaluation, and planning has also been used to assist retrieval-augmented LLMs to significantly outperform prior work [334].

Others. Metacognitive methods have also been used to improve other diverse abilities of LLMs. This includes metacognitive frameworks, training pipelines, prompting methods, and decision guidance. For instance, Lin et al. [176], Qin et al. [228], Shan et al. [248] devise frameworks incorporating metacognitive self-assessment, error detection, and learning strategies to boost LLMs' ability to identify and mitigate domain-specific risks, engage in creative planning, maintain consistency in role-playing tasks, and learn from distillation training. Toward risk mitigation, Shan et al. [248] posit that metacognitive monitoring strategies can help to navigate the generalization-specialization tradeoff by bypassing the need for task-specific training while avoiding the loss of depth associated with overly general solutions. Qin et al. [228] show that metacognitive evaluations during training can be used to provide interpretable intermediate generation states. Lin et al. [176] also find that metacognitive learning can foster creativity and improve efficiency, enabling models to perform robotic tasks with minimal demonstrations. Beyond this, the incorporation of metacognitive knowledge during training by attaching annotations of, for example, skills required to solve a question, to training samples has been shown to help reduce catastrophic forgetting [322]. Metacognitive prompting methods have been applied toward improving prompt engineering [198, 229] and models' ability to perform causal inference [218], self-identify and reduce biases [105], detect sarcasm [152], and generate ontologies [178]. Additionally, metacognitive signals (e.g., self-evaluation of competence) have been

leveraged to support adaptive decision-making in tool use and pedagogical settings, enabling models to dynamically regulate when to invoke external tools and guide human learning processes in a more interpretable manner [165, 184].

Taken together, the works discussed in this section demonstrate that metacognitive monitoring and control behaviors can lead to tangible benefits for diverse downstream abilities of LLMs. Despite this, conclusions appear scattered across models and tasks, some of which are outdated.

6.2 Improving Learning Efficacy

We introduce metacognitive methods by which to improve the efficiency and efficacy of learning and skill acquisition in LLMs.

Task Monitoring. The ability to reliably evaluate and track task progress is a fundamental aspect of efficient and effective learning. Such monitoring enables individuals to regularly compare their current state of understanding against an intended goal and adjust accordingly. However, it is often hindered in LLMs due to their poor metacognitive calibration, systematic overconfidence, and limited self-monitoring and regulation (§4.2). To this end, recent studies have begun to investigate whether explicit metacognitive supervision can serve as a solution. In the context of deep search tasks, Sun et al. [265] observe that agent failures often are not purely attributable to weak reasoning, but rather to a lack of mechanisms to detect when retrieval states and reasoning trajectories become unreliable. Thus, they introduce a hierarchical monitoring framework consisting of a fast consistency monitor and slow experience-driven monitor. The fast consistency monitor is responsible for ensuring the model's reasoning aligns with external evidence. If there is discrepancy, then the slow experience-driven monitor accesses memory records to generate corrective actions. This simulates a form of metacognitive monitoring that enables improved regulation of task completion progress. Similarly, Roy and Roy [235] present a fine-tuning approach to improve models' metacognitive calibration when learning physics-based tasks, leading to improved prediction performance and monitoring capabilities and providing further support for the value of supervisory frameworks for metacognitive monitoring. Relatedly, Fan et al. [76] present a framework emulating the learning process of humans with meta-memory mechanisms, enabling multimodal LLMs to better (metacognitively) monitor the scope and reliability of new information. This improves learning success against prior cognition-inspired approaches and further improves knowledge editing on multiple benchmarks.

Self-Critique. Reflection is a metacognitive trait that promotes error identification and revision toward improved learning outcomes. To this end, recent work has implemented metacognitive self-critique strategies to enhance the reflective capabilities and subsequent performance of LLMs. Hills [105] demonstrate that LLMs can be primed to evaluate and convey biased decisions in their reasoning via metacognitive prompting. Wang et al. [289] additionally move beyond superficial reflection behavior via a training framework that leverages high-quality reflections to construct reward signals for reinforcement learning that guide models to internalize self-correction procedures, leading to effective self-distillation. Notably, they provide evidence refuting earlier propositions that models' inherent reflection behaviors are truly metacognitive.

Strategy Selection. The ability to make accurate judgments of what method to use and when one is better than others is another crucial feature of learning and planning that is informed by metacognition [34]. Existing metacognition-inspired approaches toward this mainly focus on LLM agents and are discussed previously in §5.3. One application of this ability is toward ensemble selection, where multiple expert models are available and chosen based on a selection policy that maximizes the cumulative reward. In this direction, Trinh et al. [278] leverage metacognitive sensitivity to inform model selection, finding that this mechanism improves joint accuracy over individual performance and is superior to standard confidence-based selection—which often obscures metacognitive qualities that are important for a complete understanding of the reliability of models' confidence signals and subsequent decisions regarding strategy selection (§4.2).

Self-Improvement. More generally, the ability of a model to generalizably and iteratively improve its performance and capabilities on its own is a key aspect of metacognition that underpins efficient and effective learning [185]. A few works investigate this by proposing frameworks for models to self-improve upon their reasoning, correction, and evolution capabilities. For instance, Li and

Qiu [168] propose a framework in which models can self-improve without additional training by "pre-thinking" using high-confidence reasoning paths stored in memory, reinforcing the idea that self-improvement depends on effectively abstracting and recalling successful and relevant problem-solving. Complementing this perspective, Qu et al. [230] show that self-improvement can be implemented via self-evolving mechanisms, introducing a fine-tuning framework to enable continual introspection and self-refinement in LLMs. Bao et al. [11] further implement a framework for continual and autonomous improvement that integrates metacognitive monitoring, adaptive memory, and dynamic learning; the aim is to enable agents to identify knowledge gaps, seek relevant information, and refine their internal representations over time.

7 Applications of LLM Metacognition

Conferring LLMs with more human-like metacognitive capacities and behaviors can be useful toward enhancing downstream human-AI interactions. The study of such benefits is relatively limited; we summarize three primary directions.

Human-AI Decision-Making. Collaborative human–AI decision-making has proven to be a powerful approach for improving task outcomes, with the metacognitive performance of LLM-based and other AI systems playing a critical role in maximizing decision accuracy. It is well-established in psychology that metacognitive ratings are important for optimal joint decisions by humans [150]. In a similar vein, recent works have highlighted the complementary contribution of *metacognitive sensitivity* (recall: the ability to assign confidence scores that discriminate correct from incorrect responses) in conjunction with predictive accuracy of LLM-based and other AI systems toward optimizing hybrid decision-making outcomes [171, 170]. Such studies have formulated theoretical frameworks to model and evaluate the relative impact of these two properties in hybrid decision-making contexts, showing that better AI metacognitive sensitivity increases human decision performance and that agents with lower accuracy but higher metacognitive sensitivity can actually improve overall accuracy. These results underscore the importance of prioritizing metacognitive sensitivity of models to achieve superior decision outcomes. More broadly, the reliability of confidence estimates issued by AI systems is critical to helping users to calibrate their trust in such systems and optimize their consideration and use of AI advice [150]. Fernandes et al. [78] find that use of AI tools can help mitigate systematic overconfidence in humans but sometimes reduces the precision with which humans assess their own performance. Colombatto et al. [54] further identify differences in self-confidence and response detail when users accept or reject LLM assistance in event planning settings.

Yet as reliance on AI expands, human metacognitive skills become increasingly necessary [116]. Stronger metacognitive capacities for LLMs and other AI systems does not remove the burden of discernment from human users. In fact, use of generative AI can impose multiple metacognitive demands on human users, even as explainability helps to offload some of the work [270]. Several studies have observed increased "metacognitive laziness" in human users who depend on LLM assistance, wherein they become less actively engaged in critical thinking and learning [75, 254]. To address this, some have proposed targeted efforts to increase metacognition-based AI literacy [237, 20], including tasking users with deliberately engaging in critical thinking and rational reasoning when considering LLM-generated information [148], and undergoing metacognition training with special prompts and quizzes [263]. The timing of feedback provided to humans during AI-assisted learning can further play a role in such metacognitive judgments [210].

User Simulation. LLM metacognition can be used to simulate human users to test AI systems for downstream settings such as teaching or therapy, in instances where it is difficult or not technically possible to effectively use real humans. Borchers et al. [22] investigate the accuracy with which LLMs anticipate novice learners' success at problem-solving and reflect their metacognitive models, finding that such LLM simulators tend to generate overly coherent and confident reasoning compared to real student think-alouds. To combat this weakness, Zheng et al. [329] propose Cognitive Echo, which fine-tunes LLMs on authentic learning interaction data to allow them to generate reasoning that resembles learners' cognitive processes and strategies more closely. This enables researchers to anticipate and capture student cognitive strategies and behaviors that would otherwise go undetected. Toward identifying high-quality student agents, Li et al. [161] introduce a pipeline to generate and filter LLM simulators by first automatically generating diverse student profiles across various attributes (e.g. demographics, personality, learning-based characteristics), then running automated

scoring and evaluating a subset of them through human expert evaluation, and lastly applying graph-based score refinement across similar profiles. Besides simulating students, LLMs can also be used to simulate clients in therapy sessions. Current simulated clients unrealistically reveal and understand their thoughts and feelings, making it difficult to assess whether a model can genuinely explore a client's internal beliefs. Therefore, Kim et al. [137] present MindVoyager, which provides controllable levels of openness and metacognition that dynamically adapt to the progress of the therapy session, and show that this setup enables more realistic testing of LLM therapists.

Pedagogy. As metacognition is widely recognized as a critical driver of effective learning [81], LLMs—and AI-powered tools more broadly—have the potential to serve as metacognitive tools in educational settings, by scaffolding reflective learning practices through intelligent design [151]. Lee et al. [153] introduce PedagogicalRL-Thinking, a framework that rewards the model's thinking process and uses pedagogically-grounded prompting, yielding significant improvements in student solve rates. Other works explore interactive designs: Holub et al. [107] use multi-agent Socratic dialogue to iteratively refine metacognitive reflection questions, Holmes et al. [106] find that LLM-generated prompts designed around metacognitive objectives consistently outperform alternatives, and Tomisu et al. [276] reconceptualize AI as a "Cognitive Mirror"—a teachable novice whose deliberate ignorance forces learners to externalize and monitor their own understanding. Gatzia and Wood [94] similarly investigate how LLM-based activities can foster metacognitive development through an "Essay Analysis Assignment" in which students critically evaluate AI-generated text.

However, this promise does not always come to fruition. Boettcher [21] finds that students rarely use LLMs for metacognitive tasks—only 23% do so in engineering lab settings. Even when students do engage, the quality of interaction varies considerably: Contel and Cusi [56] observe that many students rely on ChatGPT as a passive replacement for their own reasoning rather than using it diagnostically, while the model itself frequently fails to scaffold planning and self-monitoring. Uittenhove et al. [279] report similar findings, showing that a well-prompted metacognitive chatbot sees low engagement across three educational contexts, with no link between engagement and learning outcomes. Notably, however, Contel and Cusi [56] find that students who do express metacognitive awareness and use the LLM as a diagnostic tool see increased benefit—suggesting that realizing the metacognitive potential of LLMs in education requires not just capable models but careful pedagogical design around them.

8 Reflections & Discussion

Why metacognition? Metacognition is a key aspect of intelligence concerning the processes by which we monitor and control our own cognition. Endowing language models with metacognitive capacities can enable them to become more reliable, efficient, and effective at learning, problem-solving, decision-making, and planning. This can improve the performance of LLMs on downstream tasks in a more generalized fashion, increase adaptability to user needs, and extend fundamental LLM abilities [93, 248]. It can also make LLMs more interpretable and explainable and enhance the reliability and quality of human-AI interactions.

Are LLMs really able to exhibit metacognition? So far, it is clear that LLMs are not yet capable of robust metacognition [274]. Indications that LLMs exhibit weak or limited metacognition (§4) include (1) their systematic tendency toward overconfidence and misaligned reasoning and confidence expressions [155, 36, 245], (2) their struggle with tasks involving metacognitive judgments of knowledge, performance, and resource allocation [315, 19, 117, 118, 129, 327], (3) their fragile introspection and revision abilities [316, 103, 224, 162], and (4) their susceptibility to redundant information [239]. LLMs also appear to differ from humans in their metacognitive capacities and behaviors [260]. Still, some observations suggest that LLMs may indeed exhibit metacognitive tendencies and that these are similar to humans, including the ability to improve at metacognitive skills via training or prompting interventions (§5, §6), the exhibition of metacognition-derived abilities during reasoning and reflection (§5.2), differing levels of metacognitive performance across task settings, and the seemingly disjoint relation between performance and metacognition (§4.2). Further in-depth analysis of factors such as model architecture, optimization objectives, training data and procedure is needed, along with the development of more systematic benchmarks to measure metacognitive abilities of LLMs [321].

Additionally, there is ongoing debate about whether models can actually engage in metacognitive processes rather than superficially simulating them [227]. This mirrors broader discussion of whether models exhibit genuine understanding or merely (pretend to) imitate it [218], including questions about the authenticity of LLM reasoning. Consciousness is not necessarily a prerequisite for metacognition [213, 234, 147, 273], and it is possible for metacognition to operate without an individual being aware of it, without belief in one's own consciousness, and without introspection [133, 297, 36, 128]. Metacognition can also *functionally* operate in entities such as LLMs. However, some propose that making LLMs more human-like in their metacognition may require a normative shift in tension with the goal of using them as tools [274]. It is also important to be precise in use of the term metacognition, and to be careful to not conflate it with, e.g., simple reflection processes [73].

What are challenges to designing effective methods for LLM metacognition? Metacognitive interventions for LLMs require careful design to be efficacious. There is conflicting evidence, for example, on the utility of meta-reflective prompts in different settings [231, 180, 92, 275]. Model-specific design may be important, as adapting strategies for one model to others may neglect model-specific metacognitive representations [216]. This parallels the privileged access humans have to their own metacognition [212], with the helpfulness of different metacognitive strategies and training varying across individuals [201, 298, 271, 33]. The choice of metacognitive approach and degree to which it is psychologically informed are other factors warranting consideration. Training-based methods remain limited, with much of the existing work relying on prompting, despite spanning diverse use cases. This raises questions regarding the extent to which metacognitive strategies should be designed for generalizability versus tailored to specific tasks, the level of complexity required (e.g., whether sophisticated frameworks are necessary), and the degree to which psychological principles should be incorporated. Alternate views on (AI) metacognition [132] may be useful as well; for example, Fitousi [79] reformulate metacognitive sensitivity as an information processing problem involving multiple agents, providing a different theoretical operationalization for studying confidence and metacognitive efficiency. Unreliable self-reporting and sensitivity of metacognitive estimates to confidence elicitation methodology present additional questions for analysis and design in this space. When self-reported signals are used, the distinction between prospective and retrospective judgments can be relevant [250]. Lastly, safety implications of metacognition for LLMs are noteworthy to consider as the field progresses. It may be valuable for alignment and metacognition advances to go hand in hand, along with attention to differences between human and AI metacognition, which can enhance oversight and collaboration, respectively. It is also important to be mindful of the apparent domain specificity of LLM metacognition, as a model being metacognitively efficient in one domain and poor in another could pose deployment risks.

What are risks of LLM metacognition? Recent studies have suggested that LLMs can monitor, report on, and modulate their internal states and neural activations, and exhibit behavioral self-awareness, the ability to accurately describe or predict learned behaviors without explicit supervision [162]. Investigations such as Betley et al. [17] have also found that models can be aware of backdoor attacks and other vulnerabilities. If a model is honest, such capabilities enable it to identify problematic tendencies that can arise in its behavior (e.g., training biases, data poisoning [46, 74, 285, 31]). However, this also raises safety concerns, as a dishonest model could leverage its awareness to strategically misrepresent or purposefully obfuscate its true abilities and internal processes during evaluation [98, 25]. To this end, establishing a better understanding of the extent of LLMs' metacognitive abilities can be critical. The autonomy associated with certain metacognitive capacities (e.g., self-directed planning or learning) must also be considered, and potentially subject to monitoring in safety- or privacy-critical contexts [5]. It is also of interest to speak of the interplay between human and LLM metacognition. A human user who usually has good metacognition may confidently believe they can detect when AI is misleading them and therefore scrutinize outputs less carefully relative to user with less confidence in their own evaluative skills [199]. Vigilance against the wrong threats is another risk if users are inadequately informed of a system's metacognitive abilities and associated safeguards. Some further have posited that models capable of metacognitive processing should bear greater responsibility for actions and decisions [232, 82], representing another direction for further discourse.

Despite potential risks, metacognition for LLMs can also help to improve safety and transparency in high-stakes settings by directly combating overconfidence to reduce the likelihood of systems failing to recognize when they are pursuing flawed solution paths or detrimental strategies [233].

Metacognitive deficits are often credited as the source of discrepancies between models' internal uncertainty and expressed confidence, or between their internal decision-making and verbalized reasoning processes, so improved metacognition can further improve the faithfulness of models' communication with human users, boost interpretability, and enhance our ability to detect flawed decision-making [149, 50]. Overall, the risks and benefits of metacognition require careful attention and additional consideration.

9 Future Directions

Metacognition in LLMs remains a nascent and exploratory field, and many directions for future research are possible given the broad scope of metacognitive abilities. We highlight a representative set of directions below, beginning with more general and foundational avenues before moving on to emerging topics and concluding with more long-term, speculative questions.

Measuring, Understanding, Improving, and Applying LLM Metacognition. While LLM metacognition has wide-ranging applications, more systematic, robust, and comprehensive methodologies to measure and quantify LLMs' existing metacognitive faculties are required to advance progress. Current paradigms to assess particular metacognitive abilities of LLMs or quantify their metacognitive skills in specific task settings lack consistency in measurement techniques, confidence elicitation strategies, task formulations, models, and other experimental factors such as temperature and dataset size (§4), making it difficult to assess the reliability and generalizability of resulting observations. Metacognition-focused benchmarks suffer from similar limitations in systematicity and scope and are few in number (§4.1). Tasks considered can be artificial and may not reflect real-world settings. Further, metacognitive capacities of LLMs are largely neglected in general evaluation benchmarks, despite the various ways in which they impact the downstream performance and utility of LLMs and the complementary insights metacognitive metrics can provide beyond standard measures of confidence calibration (§4.2).

It is also not yet clear how we can explain the emergence of metacognition-like processing or behaviors in LLMs, nor whether such abilities manifest in a generalized or domain-specific fashion. Understanding the role of post-training procedure, training data properties (e.g., diversity, prompt structure), model architecture, and other factors on LLM metacognition represents one line of inquiry in this direction. The application of mechanistic interpretability techniques or other targeted analysis to uncover underlying mechanisms responsible for observed metacognitive behaviors presents another unexplored and fruitful avenue.

Toward improving and applying LLM metacognition, future work could investigate the devision and use of metacognition-inspired mechanisms for a wider range of models, task scenarios, and practical applications, potentially in combination with existing approaches. For example, metacognitive reward design has been considered by several works and yielded success in improving the faithfulness and calibration of LLMs (§5.1, §6). The extent to which metacognition can support learning efficiency and efficacy and other fundamentally relevant functional processes also remains to be explored. More sensitive and well-calibrated metacognition in LLMs may further unlock new capacities such as self-directed behaviors and learning, curiosity-driven exploration, and creativity [260].

Metacognition and Creativity. Creativity is important for a variety of cognitive abilities and tasks, such as problem-solving [9], robotic planning [176], and research ideation [182]. In humans, metacognitive feelings are integral to the generation, evaluation, and selection of creative ideas [226]. It is also of importance in co-creative settings, which are an increasingly prevalent form of human–AI interaction wherein metacognition-supported reflection can mediate how users adapt creative strategies over time and interpret system behaviors and design intentions [287]. Preliminary work shows metacognitive prompts can improve the explainability and efficacy of human-AI scientific ideation systems and help researchers to examine how ideas evolve [182]. Further benefits to task performance and downstream applications may be achieved by studying creative mechanisms for LLMs based on metacognitive principles.

Metacognition and Self Improvement. As metacognition is a fundamental part of learning that enables active evaluation, reflection, and adaptation, it may be key to the development of self-improving agents. Self-improvement refers to the ability to continuously acquire new capabilities

and experience with minimal (human) supervision. Current approaches to instantiate this ability in LLM-based agentic systems are often rigid or lack generalization and scalability. It is suggested by some [185] that this is due to reliance on purely *external* metacognitive mechanisms, which depend on fixed, human-designed implementations of metacognition, instead of intrinsic metacognitive learning. Regardless, endowing LLMs with improved metacognitive capacities represents a promising direction toward more generalized and aligned self-improvement. How best to balance metacognitive responsibilities between humans and agents in this process is an open question.

Metacognition and Theory of Mind. The study of metacognition in LLMs can be helpful toward understanding their theory of mind capabilities—another aspect of human cognition of recent interest for LLMs, which has implications for collaborative or social environments involving multiple agents. Theory of mind (ToM) describes the ability to attribute unobservable mental states (e.g., intentions, desires, beliefs) to others, and it is closely related to mentalizing, which refers to the application of metacognition-like monitoring to other individuals [89, 142].[10] A number of works have suggested that LLMs can demonstrate emergent ToM abilities [6, 45, 110, 283], although findings are mixed. ToM can improve models' ability to model user psychology, learn about users in a humanistic fashion, and generate responses that are more effective, adaptive, and contextually appropriate (e.g., in terms of style, scope, tone, persuasiveness), with further utility in pedagogical and other agentic settings [215]. Notably, Leer et al. [154] demonstrate that metacognitive prompting can improve ToM in LLMs and reduce ToM prediction error in AI tutoring applications. Considering the functional similarity of metacognition and ToM, conferring LLMs with better metacognition may have implications for their ability to model and reason about the cognitive processes of other agents or human users.

The Prospect of Meta-Metacognition. Recent evidence suggests that humans possess the meta-metacognitive ability to assess the quality of their own metacognitive judgments of confidence [200, 330, 242]. Whether LLMs possess this ability has yet to be studied and presents another interesting direction for future work toward characterizing metacognition in LLMs. Comparison between human and LLM metacognition may also provide valuable insights toward this and other broader questions, and serve as a framework for more principled accounts of LLM metacognition.

10 Conclusion

This paper provides the first comprehensive and up-to-date review of the current state of research and knowledge on metacognition in LLMs. We organize and unify existing work on this topic, discussing techniques for measuring and eliciting metacognition in LLMs, methods to improve and apply LLMs' metacognitive abilities, findings and implications of work in the area, and challenges and open directions for future research. While LLMs have made remarkable progress on many NLP tasks and in diverse downstream settings, the extent to which they can display, acquire, and apply metacognitive faculties remains unclear. Further investigation is needed to understand these, as well as whether models are capable of genuine metacognition or simply simulating memorized patterns, how metacognition can facilitate other desirable behaviors and qualities, and the implications of LLM metacognition for safe oversight, deployment, and human-AI interactions. We hope this paper serves as a foundation for further exploration and discourse regarding this important topic.

Limitations

In this paper, our aim is to provide a systematic and comprehensive overview of existing research on metacognition in LLMs. While we reference numerous studies related to this topic and make the best effort to cover the most relevant works to date, it is possible that some work was overlooked. Additionally, given accelerating interest in the field, there may have been new contributions produced concurrently with writing and review of this paper. We encourage readers to also consult future papers citing this one for a more updated understanding of methods and applications of metacognition in LLMs. Finally, we note that within our discussion, where model confidence is mentioned, we primarily address it through the lens of metacognition rather than providing a comprehensive treatment of calibration, which is a well-established area with numerous existing surveys [112, 301].

[10] For example, taking into account an individual's mental states (monitoring) and using this to predict behavior (control).

Ethics Statement

As this paper did not conduct experiments, engage with sensitive datasets, or employ annotators, the work itself does not have ethical implications. Nevertheless, metacognition in LLMs is relevant to confidence calibration, self-awareness, and self-improvement, which present implications for oversight, safety, and alignment that merit careful consideration as the field advances.

References

[1] Christopher Ackerman. Evidence for limited metacognition in LLMs. *arXiv preprint arXiv:2509.21545*, 2025.

[2] Rakefet Ackerman and Valerie A Thompson. Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in cognitive sciences*, 21(8):607–617, 2017.

[3] Laksh Advani. When small models are right for wrong reasons: Process verification for trustworthy agents. *arXiv preprint arXiv:2601.00513*, 2026.

[4] Ahmet Oguz Akturk and Ismail Sahin. Literature review on metacognition and its measurement. *Procedia-Social and Behavioral Sciences*, 15:3731–3736, 2011.

[5] Linga Reddy Alva and Bishwajeet Pandey. Agentic AI systems in the age of generative models: architectures, cloud scalability, and real-world applications. *Artificial Intelligence Review*, 59 (3):88, 2026.

[6] Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. Can LLMs reason like humans? assessing theory of mind reasoning in LLMs for open-ended questions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 34–44, 2024.

[7] Michael L Anderson and Donald R Perlis. Logic, self-awareness and self-improvement: The metacognitive loop and the problem of brittleness. *Journal of Logic and Computation*, 15(1): 21–40, 2005.

[8] Thea Aviss. State stream transformer (sst): Emergent metacognitive behaviours through latent state persistence. *arXiv preprint arXiv:2501.18356*, 2025.

[9] Tian Bai, Yongwang Cao, Yan Ge, and Haitao Yu. Mp: Endowing large language models with lateral thinking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22): 23460–23468, Apr. 2025. doi: 10.1609/aaai.v39i22.34514. URL `https://ojs.aaai.org/index.php/AAAI/article/view/34514`.

[10] Ishwar B Balappanawar, Vamshi Krishna Bonagiri, Anish R Joishy, Manas Gaur, Krishnaprasad Thirunarayan, and Ponnurangam Kumaraguru. If pigs could fly... can LLMs logically reason through counterfactuals? *arXiv preprint arXiv:2505.22318*, 2025.

[11] Chongyu Bao, Ruimin Dai, Yangbo Shen, Runyang Jian, Jinghan Zhang, Xiaolan Liu, and Kunpeng Liu. Galaxy: A cognition-centered framework for proactive, privacy-preserving, and self-evolving LLM agents. *arXiv preprint arXiv:2508.03991*, 2025.

[12] Casey O Barkan, Sid Black, and Oliver Sourbut. Do large language models know what they are capable of? *arXiv preprint arXiv:2512.24661*, 2025.

[13] Adam B Barrett, Zoltan Dienes, and Anil K Seth. Measures of metacognition on signal-detection theoretic models. *Psychological methods*, 18(4):535, 2013.

[14] Evan Becker and Stefano Soatto. Cycles of thought: Measuring LLM confidence through stable explanations. *arXiv preprint arXiv:2406.03441*, 2024.

[15] Gasper Begus, Maksymilian Dabkowski, and Ryan Rhodes. Large linguistic models: Investigating LLMs' metalinguistic abilities. *IEEE Transactions on Artificial Intelligence*, 2025.

[16] M Bergamaschi Ganapini, M Campbell, F Fabiano, L Horesh, J Lenchner, A Loreggia, N Mattei, F Rossi, B Srivastava, and KB Venable. Fast, slow, and metacognitive thinking in AI. *npj Artificial Intelligence*, 1(1):27, 2025.

[17] 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. *arXiv preprint arXiv:2501.11120*, 2025.

[18] Ahsan Bilal, Muhammad Ahmed Mohsin, Muhammad Umer, Muhammad Awais Khan Bangash, and Muhammad Ali Jamshed. Meta-thinking in LLMs via multi-agent reinforcement learning: A survey. *arXiv preprint arXiv:2504.14520*, 2025.

[19] 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 *International Conference on Learning Representations*, volume 2025, pages 3710–3756, 2025.

[20] Markus Bink, Marten Risius, David Elsweiler, and Udo Kruschwitz. " can you tell me?": Designing copilots to support human judgement in online information seeking. In *Proceedings of the 2026 Conference on Human Information Interaction and Retrieval*, pages 172–182, 2026.

[21] Konrad ER Boettcher. Autonomous usage of LLMs in scenario-based laboratory learning in engineering education. In *2025 Yearbook Emerging Technologies in Learning*, pages 147–161. Springer, 2026.

[22] Conrad Borchers, Jill-Jênn Vie, and Roger Azevedo. Large language models as students who think aloud: Overly coherent, verbose, and confident. *arXiv preprint arXiv:2602.01015*, 2026.

[23] Ranajoy Bose. *Agentic RAG: Autonomous Information Systems*, pages 611–691. Apress, Berkeley, CA, 2025. ISBN 979-8-8688-1808-0. doi: 10.1007/979-8-8688-1808-0_13. URL `https://doi.org/10.1007/979-8-8688-1808-0_13`.

[24] Kendrick Boyd, Kevin H Eng, and C David Page. Area under the precision-recall curve: point estimates and confidence intervals. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 451–466. Springer, 2013.

[25] Matthew Bozoukov, Matthew Nguyen, Shubhkarman Singh, Bart Bussmann, and Patrick Leask. Emergent mechanisms of self-awareness in LLMs. In *Second Workshop on XAI4Science: From Understanding Model Behavior to Discovering New Scientific Knowledge*, 2025. URL `https://openreview.net/forum?id=6GGhnrQ2EV`.

[26] Jon-Paul Cacioli. Do LLMs know what they know? measuring metacognitive efficiency with signal detection theory. *arXiv preprint arXiv:2603.25112*, 2026.

[27] Jon-Paul Cacioli. LLMs as signal detectors: Sensitivity, bias, and the temperature-criterion analogy. *arXiv preprint arXiv:2603.14893*, 2026.

[28] Jon-Paul Cacioli. The metacognitive monitoring battery: A cross-domain benchmark for llm self-monitoring. *arXiv preprint arXiv:2604.15702*, 2026.

[29] Qi Cao, Yufan Wang, Peijia Qin, Shuhao Zhang, and Pengtao Xie. Llms know when they know, but do not act on it: A metacognitive harness for test-time scaling. *arXiv preprint arXiv:2605.14186*, 2026.

[30] Herman Cappelen and Josh Dever. Introspective machines: Are LLMs better at self-reflection than humans? *Philosophical Perspectives*, 38(1):189–196, 2024.

[31] Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 407–425. IEEE, 2024.

[32] Manuel F Caro, Darsana P Josyula, Michael T Cox, and Jovani A Jiménez. Design and validation of a metamodel for metacognition support in artificial intelligent systems. *Biologically Inspired Cognitive Architectures*, 9:82–104, 2014.

[33] Jason Carpenter, Maxine T Sherman, Rogier A Kievit, Anil K Seth, Hakwan Lau, and Stephen M Fleming. Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, 148(1):51, 2019.

[34] Melanie Cary and Lynne M Reder. Metacognition in strategy selection: Giving consciousness too much credit. In *Metacognition: Process, function and use*, pages 63–77. Springer, 2002.

[35] Trent Cash and Daniel Oppenheimer. Worth the weight: Modern LLMs demonstrate accurate metacognitive knowledge of decision weights in multi-attribute choice. *PsyArXiv*, 2025. doi: 10.31234/osf.io/zq6md_v1. URL `https://doi.org/10.31234/osf.io/zq6md_v1`.

[36] Trent N Cash, Daniel M Oppenheimer, Sara Christie, and Mira Devgan. Quantifying uncert-AI-nty: Testing the accuracy of LLMs' confidence judgments. *Memory & Cognition*, 54(2): 375–400, 2026.

[37] Harrison Chase. LangChain, October 2022. URL `https://github.com/langchain-ai/langchain`.

[38] Athanasia Chatzipanteli, Vasilis Grammatikopoulos, and Athanasios Gregoriadis. Development and evaluation of metacognition in early childhood education. *Early child development and care*, 184(8):1223–1232, 2014.

[39] Hanting Chen, Yasheng Wang, Kai Han, Dong Li, Lin Li, Zhenni Bi, Jinpeng Li, Haoyu Wang, Fei Mi, Mingjian Zhu, et al. Pangu embedded: An efficient dual-system LLM reasoner with metacognition. *arXiv preprint arXiv:2505.22375*, 2025.

[40] Hao Chen, Ye He, Yuchun Fan, Yukun Yan, Zhenghao Liu, Qingfu Zhu, Maosong Sun, and Wanxiang Che. Know more, know clearer: A meta-cognitive framework for knowledge augmentation in large language models. *arXiv preprint arXiv:2602.12996*, 2026.

[41] Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.283. URL `https://aclanthology.org/2024.acl-long.283/`.

[42] Lihu Chen, Gerard de Melo, Fabian M Suchanek, and Gaël Varoquaux. Query-level uncertainty in large language models. *arXiv preprint arXiv:2506.09669*, 2025.

[43] Patricia Chen, Omar Chavez, Desmond C Ong, and Brenda Gunderson. Strategic resource use for learning: A self-administered intervention that guides self-reflection on effective resource use enhances academic performance. *Psychological science*, 28(6):774–785, 2017.

[44] Rubing Chen, Jian Wang, Wenjie Li, Xiao-Yong Wei, and Qing Li. To retrieve or to think? an agentic approach for context evolution. *arXiv preprint arXiv:2601.08747*, 2026.

[45] Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. Theory of mind in large language models: Assessment and enhancement. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31539–31558, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1522. URL `https://aclanthology.org/2025.acl-long.1522/`.

[46] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[47] Yuxin Chen, Yu Wang, Yi Zhang, Ziang Ye, Zhengzhou Cai, Yaorui Shi, Qi Gu, Hui Su, Xunliang Cai, Xiang Wang, et al. Learning to self-verify makes language models better reasoners. *arXiv preprint arXiv:2602.07594*, 2026.

[48] Siyuan Cheng, Bozhong Tian, YanChao Hao, and Zheng Wei. Prism-mcts: Learning from reasoning trajectories with metacognitive reflection. *arXiv preprint arXiv:2604.05424*, 2026.

[49] Prateek Chhikara. Mind the confidence gap: Overconfidence, calibration, and distractor effects in large language models. *arXiv preprint arXiv:2502.11028*, 2025.

[50] James Chua and Owain Evans. Are deepseek R1 and other reasoning models more faithful? *arXiv preprint arXiv:2501.08156*, 2025.

[51] Stephen Chung, Wenyu Du, and Jie Fu. Learning from failures in multi-attempt reinforcement learning. *arXiv preprint arXiv:2503.04808*, 2025.

[52] Ahsen Çini, Sanna Järvelä, Muhterem Dindar, and Jonna Malmberg. How multiple levels of metacognitive awareness operate in collaborative problem solving. *Metacognition and Learning*, 18(3):891–922, 2023.

[53] Marisa T Cohen. The importance of self-regulation for college student learning. *College Student Journal*, 46(4):892–903, 2012.

[54] Clara Colombatto, Sean Rintel, and Lev Tankelevitch. Metacognition and confidence dynamics in advice taking from generative ai. *arXiv preprint arXiv:2510.26508*, 2025.

[55] Iulia M Comsa and Murray Shanahan. Does it make sense to speak of introspection in large language models? *arXiv preprint arXiv:2506.05068*, 2025.

[56] Francesco Contel and Annalisa Cusi. Investigating the role of ChatGPT in supporting metacognitive processes during problem-solving activities. *Digital Experiences in Mathematics Education*, 11(1):167–191, 2025.

[57] Brendan Conway-Smith and Robert L. West. Toward autonomy: Metacognitive learning for enhanced AI performance. *Proceedings of the AAAI Symposium Series*, 3(1):545–546, May 2024. doi: 10.1609/aaaiss.v3i1.31270. URL `https://ojs.aaai.org/index.php/AAAI-SS/article/view/31270`.

[58] Wendi Cui, Sirui Bi, Anni Huang, Wei Wang, Damien Lopez, Fangzhou Yan, and Jiaxin Zhang. Thinking about thinking: Metacognitive influence tracing for reliable LLM reasoning, 2025. URL `https://openreview.net/forum?id=aJShPeHb5z`.

[59] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models, 2024.

[60] Yuyang Dai. Rescaling confidence: What scale design reveals about LLM metacognition. *arXiv preprint arXiv:2603.09309*, 2026.

[61] Peter Dayan. Metacognitive information theory. *Open Mind*, 7:392–411, 2023.

[62] Aniket Didolkar, Anirudh Goyal, Nan R Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. Metacognitive capabilities of LLMs: An exploration in mathematical problem solving. *Advances in Neural Information Processing Systems*, 37:19783–19812, 2024.

[63] Aniket Didolkar, Nicolas Ballas, Sanjeev Arora, and Anirudh Goyal. Metacognitive reuse: Turning recurring LLM reasoning into concise behaviors. *arXiv preprint arXiv:2509.13237*, 2025.

[64] Haonan Dong, Haoran Ye, Wenhao Zhu, Kehan Jiang, and Guojie Song. Meta-R1: Empowering large reasoning models with metacognition. *arXiv preprint arXiv:2508.17291*, 2025.

[65] Jianshuo Dong, Yujia Fu, Chuanrui Hu, Chao Zhang, and Han Qiu. Towards understanding the cognitive habits of large reasoning models. *arXiv preprint arXiv:2506.21571*, 2025.

[66] Yanrui Du, Yibo Gao, Sendong Zhao, Jiayun Li, Haochun Wang, Qika Lin, Kai He, Bing Qin, and Mengling Feng. From latent signals to reflection behavior: Tracing meta-cognitive activation trajectory in R1-style LLMs. *arXiv preprint arXiv:2602.01999*, 2026.

[67] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.276. URL https://aclanthology.org/2024.acl-long.276/.

[68] John Dunlosky and Robert A Bjork. *Handbook of metamemory and memory*. Psychology Press, 2013.

[69] John Dunlosky and Janet Metcalfe. *Metacognition*. Sage Publications, 2008.

[70] David Dunning. The dunning–kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology*, volume 44, pages 247–296. Elsevier, 2011.

[71] Roy Eisenstadt, Itamar Zimerman, and Lior Wolf. Overclocking LLM reasoning: Monitoring and controlling thinking path lengths in LLMs. *arXiv preprint arXiv:2506.07240*, 2025.

[72] Abraham Paul Elenjical, Vivek Hruday Kavuri, and Vasudeva Varma. Think^2: Grounded metacognitive reasoning in large language models. *arXiv preprint arXiv:2602.18806*, 2026.

[73] Ryan Erbe. Consciousness and ai: A meta-reflective framework. *Available at SSRN 5854902*, 2025.

[74] Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. Truthful AI: Developing and governing AI that does not lie. *arXiv preprint arXiv:2110.06674*, 2021.

[75] Yizhou Fan, Luzhen Tang, Huixiao Le, Kejie Shen, Shufang Tan, Yueying Zhao, Yuan Shen, Xinyu Li, and Dragan Gašević. Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56(2):489–530, 2025.

[76] Zhaoyu Fan, Kaihang Pan, Mingze Zhou, Bosheng Qin, Juncheng Li, Shengyu Zhang, Wenqiao Zhang, Siliang Tang, Fei Wu, and Yueting Zhuang. Towards meta-cognitive knowledge editing for multimodal LLMs. In *Proceedings of the ACM Web Conference 2026*, pages 4137–4148, 2026.

[77] Jiing Fang and Wei Chen. Probabilistic chain-of-evidence: Enhancing factual accuracy and uncertainty reasoning in large language models via prompt engineering. 2026.

[78] Daniela Fernandes, Steeven Villa, Salla Nicholls, Otso Haavisto, Daniel Buschek, Albrecht Schmidt, Thomas Kosch, Chenxinran Shen, and Robin Welsch. AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior*, page 108779, 2025.

[79] Daniel Fitousi. Bits of confidence: Metacognition as uncertainty reduction. *Psychonomic Bulletin & Review*, 32(6):2734–2762, 2025.

[80] Lisa M Fitzgerald, Mahnaz Arvaneh, and Paul M Dockree. Domain-specific and domain-general processes underlying metacognitive judgments. *Consciousness and Cognition*, 49: 264–277, 2017.

[81] John H Flavell. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906, 1979.

[82] Brendan Fleig-Goldstein. Meta-reflective capacities, normative commitments, and responsible ai. 2025.

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

[84] Stephen M Fleming and Nathaniel D Daw. Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological review*, 124(1):91, 2017.

[85] Stephen M Fleming and Raymond J Dolan. The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1338–1349, 2012.

[86] Stephen M Fleming and Hakwan C Lau. How to measure metacognition. *Frontiers in human neuroscience*, 8:82285, 2014.

[87] Stephen M Fleming, Raymond J Dolan, and Christopher D Frith. Metacognition: computation, biology and function. *Philosophical transactions of the royal society B: Biological sciences*, 367(1594):1280–1286, 2012.

[88] Damien S Fleur, Bert Bredeweg, and Wouter van den Bos. Metacognition: ideas and insights from neuro-and educational sciences. *npj Science of Learning*, 6(1):13, 2021.

[89] 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.

[90] Zhizhang Fu, Yuancheng Gu, Chenkai Hu, Hanmeng Liu, and Yue Zhang. Reefbench: Quantifying the reasoning efficiency of LLMs. *arXiv preprint arXiv:2601.03550*, 2026.

[91] Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

[92] Areeb Gani, Asal Meskin, Gabrielle Kaili-May Liu, and Arman Cohan. Quantifying faithful confidence expression in large reasoning models, 2026. URL `https://arxiv.org/abs/2606.03969`.

[93] Peizhong Gao, Ao Xie, Shaoguang Mao, Wenshan Wu, Yan Xia, Haipeng Mi, and Furu Wei. Meta reasoning for large language models. *arXiv preprint arXiv:2406.11698*, 2024.

[94] Dimitria Electra Gatzia and Moriah K Wood. Developing metacognition through LLM-enhanced writing assignments: Critical thinking exercises for scientific writing. *The American Biology Teacher*, 88(1):12–21, 2026.

[95] Loris Gaven, Thomas Carta, Clément Romac, Cédric Colas, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. Magellan: Metacognitive predictions of learning progress guide autotelic LLM agents in large goal spaces. *arXiv preprint arXiv:2502.07709*, 2025.

[96] Tomer Geva, Ariel Goldstein, Eran Lary, and Coral Levy. Do LLMs exhibit human-like cognitive biases? a large-scale systematic evaluation. *A Large-Scale Systematic Evaluation (September 17, 2025)*, 2025.

[97] W Brier Glenn et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[98] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

[99] Yashvir S. Grewal, Edwin V. Bonilla, and Thang D. Bui. Improving uncertainty quantification in large language models via semantic embeddings. *arXiv preprint arXiv:2410.22685*, 2024.

[100] 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.

[101] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[102] Rui Ha, Chaozhuo Li, Rui Pu, and Sen Su. From" aha moments" to controllable thinking: Toward meta-cognitive reasoning in large reasoning models via decoupled reasoning and control. *arXiv preprint arXiv:2508.04460*, 2025.

[103] Ely Hahami, Lavik Jain, and Ishaan Sinha. Feeling the strength but not the source: Partial introspection in LLMs. *arXiv preprint arXiv:2512.12411*, 2025.

[104] Ahshanul Haque. Meta-of-thought: Reasoning about reasoning in large language models. *Authorea Preprints*, 2025.

[105] Thomas T Hills. Could you be wrong: Metacognitive prompts for improving human decision making help LLMs identify their own biases. *AI*, 7(1):33, 2026.

[106] Langdon Holmes, Adam Coscia, Scott Crossley, Joon Suh Choi, and Wesley Morris. LLM prompt evaluation for educational applications. *arXiv preprint arXiv:2601.16134*, 2026.

[107] Ondřej Holub, Essi Ryymin, and Rodrigo Alves. Reflecting in the reflection: Integrating a socratic questioning framework into automated ai-based question generation. *arXiv preprint arXiv:2601.14798*, 2026.

[108] Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling. *arXiv preprint arXiv:2311.08718*, 2024.

[109] Xinmeng Hou, Peiliang Gong, Bohao Qu, Wuqi Wang, Qing Guo, and Yang Liu. Learn like humans: Use meta-cognitive reflection for efficient self-improvement. *arXiv preprint arXiv:2601.11974*, 2026.

[110] Jennifer Hu, Felix Sosa, and Tomer Ullman. Re-evaluating theory of mind evaluation in large language models. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 380(1932):20230499, 08 2025. ISSN 0962-8436. doi: 10.1098/rstb.2023.0499. URL `https://doi.org/10.1098/rstb.2023.0499`.

[111] Fan Huang, Haewoon Kwak, and Jisun An. Vulnerability of LLMs' belief systems? LLMs belief resistance check through strategic persuasive conversation interventions. *arXiv preprint arXiv:2601.13590*, 2026.

[112] Hsiu-Yuan Huang, Yutong Yang, Zhaoxi Zhang, Sanwoo Lee, and Yunfang Wu. A survey of uncertainty estimation in LLMs: Theory meets practice. *arXiv preprint arXiv:2410.15326*, 2024.

[113] Xinyue Huang, Zeyu Wang, Xin Liu, Yueqi Tian, and Qian Leng. Towards interpretable and consistent multi-step mathematical reasoning in large language models. In *2025 5th International Symposium on Artificial Intelligence and Intelligent Manufacturing (AIIM)*, pages 343–346. IEEE, 2025.

[114] Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. Look before you leap: An exploratory study of uncertainty analysis for large language models. *IEEE Transactions on Software Engineering*, 51(2):413–429, February 2025. ISSN 2326-3881. doi: 10.1109/tse.2024.3519464. URL `http://dx.doi.org/10.1109/TSE.2024.3519464`.

[115] Thomas Huber and Christina Niklaus. LLMs meet bloom's taxonomy: A cognitive view on large language model evaluations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5211–5246, 2025.

[116] Markus Huff and Elanur Ulakçı. Metacognitive monitoring: A human ability beyond generative artificial intelligence. *arXiv preprint arXiv:2410.13392*, pages arXiv–2410, 2024.

[117] Markus Huff and Elanur Ulakci. Judgments of learning distinguish humans from large language models in predicting memory. *Scientific Reports*, 15(1):35030, 2025.

[118] Seonjeong Hwang, Hyounghun Kim, and Gary Geunbae Lee. Can LLMs estimate cognitive complexity of reading comprehension items? *arXiv preprint arXiv:2510.25064*, 2025.

[119] Sajid Iqbal. Do AI know what they know? exploring metacognition in LLMs. *Intelligent Data Analysis*, page 1088467X261436903, 2023.

[120] Randy Isaacson and Frank Fujita. Metacognitive knowledge monitoring and self-regulated learning. *Journal of the Scholarship of Teaching and Learning*, pages 39–55, 2006.

[121] Basab Jha, Firoj Paudel, Ujjwal Puri, Zhang Yuting, Choi Donghyuk, and Wang Junhao. Thinking about thinking: Sage-nano's inverse reasoning for self-aware language models. *arXiv preprint arXiv:2507.00092*, 2025.

[122] Mingjian Jiang, Yangjun Ruan, Sicong Huang, Saifei Liao, Silviu Pitis, Roger Baker Grosse, and Jimmy Ba. Calibrating language models via augmented prompt ensembles. In *ICML 2023 Workshop on Challenges in Deployable Generative AI*, 2023. URL `https://openreview.net/forum?id=L0dc4wqbNs`.

[123] Douglas B. Johnson, Rachel S Goodman, J. Randall Patrinely, Cosby A Stone, Eli Zimmerman, Rebecca Rigel Donald, Sam S Chang, Sean T Berkowitz, Avni P Finn, Eiman Jahangir, Elizabeth A Scoville, Tyler Reese, Debra E. Friedman, Julie A. Bastarache, Yuri F van der Heijden, Jordan Wright, Nicholas Carter, Matthew R Alexander, Jennifer H Choe, Cody A Chastain, John Zic, Sara N Horst, Isik Turker, Rajiv Agarwal, Evan C. Osmundson, Kamran Idrees, Colleen M. Kiernan, Chandrasekhar Padmanabhan, Christina Edwards Bailey, Cameron Schlegel, Lola B. Chambless, Mike Gibson, Travis J. Osterman, and Lee E. Wheless. Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the Chat-GPT model. *Research Square*, 2023. URL `https://api.semanticscholar.org/CorpusID:257437276`.

[124] Samuel GB Johnson, Amir-Hossein Karimi, Yoshua Bengio, Nick Chater, Tobias Gerstenberg, Kate Larson, Sydney Levine, Melanie Mitchell, Iyad Rahwan, Bernhard Schölkopf, et al. Imagining and building wise machines: The centrality of AI metacognition. *Trends in Cognitive Sciences*, 2025.

[125] Darsana P Josyula, Harish Vadali, Bette J Donahue, and Franklin C Hughes. Modeling metacognition for learning in artificial systems. In *2009 World Congress on Nature & Biologically Inspired Computing (NaBIC)*, pages 1419–1424. IEEE, 2009.

[126] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[127] Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

[128] Caspar Kaiser and Sean Enderby. No reliable evidence of self-reported sentience in small large language models. *arXiv preprint arXiv:2601.15334*, 2026.

[129] Sahil Kale and Vijaykant Nadadur. Line of duty: Evaluating LLM self-knowledge via consistency in feasibility boundaries. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 127–140, 2025.

[130] Priyanka Kargupta, Shuyue Stella Li, Haocheng Wang, Jinu Lee, Shan Chen, Orevaoghene Ahia, Dean Light, Thomas L Griffiths, Max Kleiman-Weiner, Jiawei Han, et al. Cognitive foundations for reasoning and their manifestation in LLMs. *arXiv preprint arXiv:2511.16660*, 2025.

[131] Ramneet Kaur, Colin Samplawski, Adam D. Cobb, Anirban Roy, Brian Matejek, Manoj Acharya, Daniel Elenius, Alexander Michael Berenbeim, John A. Pavlik, Nathaniel D. Bastian, and Susmit Jha. Addressing uncertainty in LLMs to enhance reliability in generative AI. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL `https://openreview.net/forum?id=Z3DS4Pcxct`.

[132] Mitsuo Kawato and Aurelio Cortese. From internal models toward metacognitive ai. *Biological cybernetics*, 115(5):415–430, 2021.

[133] Robert Kentridge. Metacognition and awareness. *Consciousness and cognition*, 2000.

[134] Vedant Khandelwal, Francesca Rossi, Keerthiram Murugesan, Erik Miehling, Murray Campbell, Karthikeyan Natesan Ramamurthy, and Lior Horesh. Language models coupled with metacognition can outperform reasoning models. *arXiv preprint arXiv:2508.17959*, 2025.

[135] Hyunjun Kim, Junwoo Ha, Sangyoon Yu, and Haon Park. Objexmt: Objective extraction and metacognitive calibration for LLM-as-a-judge under multi-turn jailbreaks. *arXiv preprint arXiv:2508.16889*, 2025.

[136] Junsol Kim, Shiyang Lai, Nino Scherrer, James Evans, et al. Reasoning models generate societies of thought. *arXiv preprint arXiv:2601.10825*, 2026.

[137] Minju Kim, Dongje Yoo, Yeonjun Hwang, Minseok Kang, Namyoung Kim, Minju Gwak, Beong-woo Kwak, Hyungjoo Chae, Harim Kim, Yunjoong Lee, et al. Can you share your story? modeling clients' metacognition and openness for LLM therapist evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25943–25962, 2025.

[138] Yoonjeon Kim, Doohyuk Jang, and Eunho Yang. Meta-awareness enhances reasoning models: Self-alignment reinforcement learning. *arXiv preprint arXiv:2510.03259*, 2025.

[139] Asher Koriat. Metacognition: Decision making processes in self-monitoring and self-regulation. *The Wiley Blackwell handbook of judgment and decision making*, 2:356–379, 2015.

[140] Asher Koriat and Tore Helstrup. Metacognitive aspects of memory. In *Everyday memory*, pages 251–274. Psychology Press, 2007.

[141] Asher Koriat and Ravit Levy-Sadot. Conscious and unconscious metacognition: A rejoinder, 2000.

[142] Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4(169):2, 2023.

[143] Bracha Kramarski and Zemira R Mevarech. Enhancing mathematical reasoning in the classroom: The effects of cooperative learning and metacognitive training. *American educational research journal*, 40(1):281–310, 2003.

[144] Justin Kruger and David Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121, 1999.

[145] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=VD-AYtP0dve`.

[146] Dharshan Kumaran, Nathaniel Daw, Simon Osindero, Petar Veličković, and Viorica Patraucean. Causal evidence that language models use confidence to drive behavior. *arXiv preprint arXiv:2603.22161*, 2026.

[147] Craig Kunimoto, Jeff Miller, and Harold Pashler. Confidence and accuracy of near-threshold discrimination responses. *Consciousness and cognition*, 10(3):294–340, 2001.

[148] Ted Ladd and Priyanka Shrivastava. Skeptical intelligence: Refining critical thinking for ai-powered innovation. *Available at SSRN 6060134*, 2026.

[149] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

[150] 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.

[151] HaeJin Lee, Frank Stinar, Ruohan Zong, Hannah Valdiviejas, Dong Wang, and Nigel Bosch. learning behaviors mediate the effect of ai-powered support for metacognitive calibration on learning outcomes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2025.

[152] Joshua Lee, Wyatt Fong, Alexander Le, Sur Shah, Kevin Han, and Kevin Zhu. Pragmatic metacognitive prompting improves LLM performance on sarcasm detection. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 63–70, 2025.

[153] Unggi Lee, Jiyeong Bae, Jaehyeon Park, Haeun Park, Taejun Park, Younghoon Jeon, Sungmin Cho, Junbo Koh, Yeil Jeong, and Gyeonggeon Lee. Rewarding how models think pedagogically: Integrating pedagogical reasoning and thinking rewards for LLMs in education. *arXiv preprint arXiv:2601.14560*, 2026.

[154] Courtland Leer, Vincent Trost, and Vineeth Voruganti. Violation of expectation via metacognitive prompting reduces theory of mind prediction error in large language models. *arXiv preprint arXiv:2310.06983*, 2023.

[155] Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in LLMs: Reward calibration in RLHF. In *International Conference on Learning Representations*, volume 2025, pages 16484–16517, 2025.

[156] Chak Tou Leong, Dingwei Chen, Heming Xia, Qingyu Yin, Sunbowen Lee, Jian Wang, and Wenjie Li. Finding relief: Shaping reasoning behavior without reasoning supervision via belief engineering. *arXiv preprint arXiv:2601.13752*, 2026.

[157] Baoxue Li and Chunhui Zhao. Self-reflection enhances large language models towards substantial academic response. *npj Artificial Intelligence*, 1(1):42, 2025.

[158] Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, Wuyue Wang, Yiqun Liu, and Minlie Huang. LegalAgentBench: Evaluating LLM agents in legal domain. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2322–2344, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.116. URL https://aclanthology.org/2025.acl-long.116/.

[159] Haoxi Li, Sikai Bai, Jie Zhang, and Song Guo. Core: Enhancing metacognition with label-free self-evaluation in lrms. *arXiv preprint arXiv:2507.06087*, 2025.

[160] Haoxi Li, Yongjiang Liu, Jie ZHANG, Shuang Chen, and Song Guo. Towards understanding metacognition in large reasoning models, 2026. URL https://openreview.net/forum?id=JGG9EdHyZc.

[161] Haoxuan Li, Jifan Yu, Xin Cong, Yang Dang, Daniel Zhang-Li, Lu Mi, Yisi Zhan, Huiqin Liu, and Zhiyuan Liu. Which type of students can LLMs act? investigating authentic simulation with graph-based human-AI collaborative system. *arXiv preprint arXiv:2502.11678*, 2025.

[162] Ji-An Li, Huadong Xiong, Robert Wilson, Marcelo G Mattar, and Marcus K Benna. Language models are capable of metacognitive monitoring and control of their internal activations. *Advances in Neural Information Processing Systems*, 38:60073–60108, 2026.

[163] Jiaqi Li, Xinyi Dong, Yang Liu, Zhizhuo Yang, Quansen Wang, Xiaobo Wang, Song-Chun Zhu, Zixia Jia, and Zilong Zheng. Reflectevo: Improving meta introspection of small LLMs by learning self-reflection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16948–16966, 2025.

[164] Weitao Li, Hao Zhou, Xuanyu Lei, Fandong Meng, Yuanhang Liu, Jingyi Ren, Ante Wang, Xiaolong Wang, Yuanchi Zhang, Fuwen Luo, et al. Enhancing llm metacognition via cognitive pairwise training. *arXiv preprint arXiv:2606.00869*, 2026.

[165] Wenjun Li, Dexun Li, Kuicai Dong, Cong Zhang, Hao Zhang, Weiwen Liu, Yasheng Wang, Ruiming Tang, and Yong Liu. Adaptive tool use in large language models with meta-cognition trigger. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13346–13370, 2025.

[166] Xiang Li. Teach large language models the concept of meta-cognition to reduce hallucination text generation, 2024. URL `https://openreview.net/forum?id=pPvK2e8o8M`.

[167] Xiaojian Li, Haoyuan Shi, Rongwu Xu, and Wei Xu. AI awareness. *arXiv preprint arXiv:2504.20084*, 2025.

[168] Xiaonan Li and Xipeng Qiu. Mot: Memory-of-thought enables ChatGPT to self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6354–6374, 2023.

[169] Yang Li, Zhiyuan He, Yuxuan Huang, Zhuhanling Xiao, Chao Yu, Meng Fang, Kun Shao, and Jun Wang. Adapting like humans: A metacognitive agent with test-time reasoning. *arXiv preprint arXiv:2511.23262*, 2025.

[170] Zhaobin Li and Mark Steyvers. The importance of metacognitive sensitivity in human-AI decision-making. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, 2025.

[171] Zhaobin Li and Mark Steyvers. Modeling the joint impact of human and AI metacognitive sensitivity on human-ai collaboration. *Journal of Mathematical Psychology*, 129:102988, 2026.

[172] Zhuoqun Li, Hongyu Lin, Yaojie Lu, Hao Xiang, Xianpei Han, and Le Sun. Meta-cognitive analysis: Evaluating declarative and procedural knowledge in datasets and large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11222–11228, 2024.

[173] Bin Liang, Cai Ke, Runcong Zhao, Qinglin Zhu, Lin Gui, Yue Yu, Hui Wang, Ruifeng Xu, and Kam-Fai Wong. Meta-memory for large language models. *IEEE Transactions on Audio, Speech and Language Processing*, 2026.

[174] Sirui Liang, Pengfei Cao, Jian Zhao, Wenhao Teng, Xiangwen Liao, Jun Zhao, and Kang Liu. Learning how to remember: A meta-cognitive management method for structured and transferable agent memory. *arXiv preprint arXiv:2601.07470*, 2026.

[175] Dean Light, Michael Theologitis, Kshitish Ghate, Shuyue Stella Li, Benjamin Newman, Chirag Shah, Aylin Caliskan, Pang Wei Koh, Dan Suciu, and Yulia Tsvetkov. Deep reasoning in general purpose agents via structured meta-cognition. *arXiv preprint arXiv:2605.11388*, 2026.

[176] Wenjie Lin, Jin Wei-Kocsis, Jiansong Zhang, Byung-Cheol Min, Dongming Gan, Paul Asunda, and Ragu Athinarayanan. Think, reflect, create: Metacognitive learning for zero-shot robotic planning with LLMs. *arXiv preprint arXiv:2505.14899*, 2025.

[177] Jack Lindsey. Emergent introspective awareness in large language models. *arXiv preprint arXiv:2601.01828*, 2026.

[178] Anna Sofia Lippolis, Miguel Ceriani, Sara Zuppiroli, and Andrea Giovanni Nuzzolese. Ontogenia: Ontology generation with metacognitive prompting in large language models. In *European Semantic Web Conference*, pages 259–265. Springer, 2024.

[179] Gabrielle Kaili-May Liu and Arman Cohan. Can llms use linguistic uncertainty markers to reliably reflect intrinsic confidence? *arXiv preprint arXiv:2605.28778*, 2026. URL `https://arxiv.org/abs/2605.28778`.

[180] Gabrielle Kaili-May Liu, Gal Yona, Avi Caciularu, Idan Szpektor, Tim GJ Rudner, and Arman Cohan. Metafaith: Faithful natural language uncertainty expression in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29600–29644, 2025.

[181] Gabrielle Kaili-May Liu, Avi Caciularu, Gal Yona, Idan Szpektor, and Arman Cohan. Reinforcement learning with metacognitive feedback elicits faithful uncertainty expression in llms, 2026. URL `https://arxiv.org/abs/2606.32032`.

[182] Houjiang Liu, Yujin Choi, Sanjana Gautam, Gabriel Jaffe, Soo Young Rieh, and Matthew Lease. Who owns creativity and who does the work? trade-offs in LLM-supported research ideation. *arXiv preprint arXiv:2601.12152*, 2026.

[183] Liping Liu, Chunhong Zhang, Likang Wu, Chuang Zhao, Zheng Hu, Ming He, and Jianping Fan. Instruct-of-reflection: Enhancing large language models iterative reflection capabilities via dynamic-meta instruction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9956–9978, 2025.

[184] Naiming Liu, Richard Baraniuk, and Shashank Sonkar. Metaclass: Metacognitive coaching for learning with adaptive self-regulation support. *arXiv preprint arXiv:2602.02457*, 2026.

[185] Tennison Liu and Mihaela Van Der Schaar. Position: Truly self-improving agents require intrinsic metacognitive learning. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

[186] Zichen Liu, Changyu Chen, Wenjun Li, Tianyu Pang, Chao Du, and Min Lin. There may not be aha moment in R1-Zero-like training—a pilot study. *https://oatllm. notion. site/oat-zero*, 2025.

[187] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

[188] Carmelo Fabio Longo, Misael Mongiovi, Luana Bulla, and Antonio Lieto. Eliciting meta-knowledge in large language models. *Cognitive Systems Research*, 91:101352, 2025.

[189] Haolang Lu, Yilian Liu, Jingxin Xu, Guoshun Nan, Yuanlong Yu, Zhican Chen, and Kun Wang. Auditing meta-cognitive hallucinations in reasoning large language models. *Advances in Neural Information Processing Systems*, 38:162500–162543, 2026.

[190] Jiaxuan Lu, Ziyu Kong, Yemin Wang, Rong Fu, Haiyuan Wan, Cheng Yang, Wenjie Lou, Haoran Sun, Lilong Wang, Yankai Jiang, et al. Beyond static tools: Test-time tool evolution for scientific reasoning. *arXiv preprint arXiv:2601.07641*, 2026.

[191] Bo Lv, Nayu Liu, Yang Shen, Xin Liu, Ping Luo, and Yue Yu. Whether LLMs know if they know: Identifying knowledge boundaries via debiased historical in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19516–19528, 2025.

[192] Ziyang Ma, Qingyue Yuan, Zhenglin Wang, and Deyu Zhou. Large language models have intrinsic meta-cognition, but need a good lens. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3460–3477, 2025.

[193] Yuxiang Mai, Qiyue Yin, Wancheng Ni, Jianwei Guo, Xiaogang Ouyang, Pei Xu, and Kaiqi Huang. Refrea: Reference-guided reasoning with meta-cognition for accurate language model agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32465–32473, 2026.

[194] Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL `https://aclanthology.org/2023.emnlp-main.557/`.

[195] Brian Maniscalco and Hakwan Lau. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition*, 21(1):422–430, 2012.

[196] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[197] Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, et al. Deepseek-R1 thoughtology: Let's think about LLM reasoning. *arXiv preprint arXiv:2504.07128*, 2025.

[198] Ian Maxwell. Meta-cognitive prompting: A comparative framework for prompt engineering in large language models. 2025. doi: 10.13140/RG.2.2.22405.46562.

[199] Andrew D Maynard. The AI cognitive trojan horse: How large language models may bypass human epistemic vigilance. *arXiv preprint arXiv:2601.07085*, 2026.

[200] Audrey Mazancieux, Michael Pereira, Nathan Faivre, Pascal Mamassian, Chris Moulin, and Souchay Céline. Towards a common conceptual space for metacognition in perception and memory. *Nature Reviews Psychology*, 2, 11 2023. doi: 10.1038/s44159-023-00245-1.

[201] Barbara L McCombs. Motivational skills training: Combining metacognitive, cognitive, and affective learning strategies. In *Learning and study strategies*, pages 141–169. Elsevier, 1988.

[202] Christine B McCormick. Metacognition and learning. *Handbook of psychology*, pages 79–102, 2003.

[203] Clara Meister, Gian Wiher, Tiago Pimentel, and Ryan Cotterell. On the probability–quality paradox in language generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 36–45, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.5. URL `https://aclanthology.org/2022.acl-short.5/`.

[204] Janet Metcalfe. Feeling of knowing in memory and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(2):288, 1986.

[205] Sascha Meyen, Frieder Göppert, Carina Schrenk, Ulrike von Luxburg, and Volker H Franz. Information-theoretic measures of metacognition: Bounds and relation to group performance. *Open Mind*, 9:1728–1762, 2025.

[206] Brian Miki and Nicholas Vincent. Step-by-step reasoning with meta cognitive prompts to reduce contextual hallucination. In *Proceedings of the CHI 2025 HEAL Workshop*, Yokohama, Japan, 2025. URL `https://heal-workshop.github.io/chi2025_papers/8_Step_By_Step_Reasoning_with_.pdf`.

[207] Jorge Morales, Hakwan Lau, and Stephen M Fleming. Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *Journal of Neuroscience*, 38(14):3534–3546, 2018.

[208] David Moshman. Metacognitive theories revisited. *Educational Psychology Review*, 30(2): 599–606, 2018.

[209] Emanuele Musumeci, Vincenzo Suriani, and Daniele Nardi. Robodata: Toward trustable question answering over ontologies through metacognitive agentic epistemology. In *Proceedings of the 5th Wikidata Workshop at ISWC 2025*, 2025. URL `https://wikidataworkshop.github.io/2025/papers/paper10.pdf`.

[210] Asikaer Nadila, Sylvain Sénécal, and Constantinos K Coursaris. *Impact of Feedback Timing on Metacognition*. PhD thesis, HEC Montréal, 2024.

[211] Atharv Naphade, Samarth Bhargav, Sean Lim, and Mcnair Shah. Me, myself, and π: Evaluating and explaining llm introspection. *arXiv preprint arXiv:2603.20276*, 2026.

[212] Thomas O Nelson. Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation*, volume 26, pages 125–173. Elsevier, 1990.

[213] Natika Newton. Metacognition and consciousness. *Pragmatics & Cognition*, 3(2):285–297, 1995.

[214] Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. *arXiv preprint arXiv:2405.20003*, 2024.

[215] Xuefei Ning, Zifu Wang, Shiyao Li, Zinan Lin, Peiran Yao, Tianyu Fu, Matthew B. Blaschko, Guohao Dai, Huazhong Yang, and Yu Wang. Can LLMs learn by teaching for better reasoning? a preliminary study. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 71188–71239. Curran Associates, Inc., 2024. doi: 10.52202/079017-2275. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/8340b085045cf13f1f0b6c2c4cc0a89c-Paper-Conference.pdf`.

[216] Nick Oh. Before you <think>, monitor: Implementing flavell's metacognitive framework in LLMs. *arXiv preprint arXiv:2510.16374*, 2025.

[217] Nick Oh and Fernand Gobet. Monitor-generate-verify (mgv): Formalising metacognitive theory for language model reasoning. *arXiv preprint arXiv:2511.04341*, 2025.

[218] Ryusei Ohtani, Yuko Sakurai, and Satoshi Oyama. Does metacognitive prompting improve causal inference in large language models? In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 458–459. IEEE, 2024.

[219] Paul Pajo. System prompt learning in large language models: Cross-disciplinary parallels with human cognition. 2025. doi: 10.13140/RG.2.2.33480.23049.

[220] Sangjun Park, Elliot Meyerson, Xin Qiu, and Risto Miikkulainen. Fine-tuning language models to know what they know. *arXiv preprint arXiv:2602.02605*, 2026.

[221] Lucia Passaro, Simone Marzeddu, Andrea Cossu, and Davide Bacciu. Awareness in LLMs improves through collaboration. In *EurIPS 2025 Workshop on Metacognition in Generative AI*, 2025.

[222] Jelena Pavlovic, Jugoslav Krstic, Luka Mitrovic, Djordje Babic, Adrijana Milosavljevic, Milena Nikolic, Tijana Karaklic, and Tijana Mitrovic. Generative AI as a metacognitive agent: a comparative mixed-method study with human participants on ICF-mimicking exam performance. *arXiv preprint arXiv:2405.05285*, 2024.

[223] Dillon Plunkett, Adam Morris, Keerthi Reddy, and Jorge Morales. Self-interpretability: LLMs can describe complex internal processes that drive their decisions. *arXiv preprint arXiv:2505.17120*, 2025.

[224] Pradyumna Shyama Prasad and Minh Nhat Nguyen. When two LLMs debate, both think they'll win. *arXiv preprint arXiv:2505.19184*, 2025.

[225] Gabriele Prato, Jerry Huang, Prasanna Parthasarathi, Shagun Sodhani, and Sarath Chandar. Do large language models know how much they know? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6054–6070, 2024.

[226] Rogelio Puente-Díaz. Metacognitive feelings as a source of information for the creative process: A conceptual exploration. *Journal of Intelligence*, 11(3):49, 2023.

[227] Ming Qian and Luyi Yang. Cognitive reasoning in translation: Evaluating chain-of-thought, explaining, metacognition, and critique in humans and general-purpose vs. advanced-reasoning large language models. In *International Conference on Human-Computer Interaction*, pages 420–437. Springer, 2025.

[228] Haiming Qin, Jiwei Zhang, Wei Zhang, KeZhong Lu, Mingyang Zhou, Hao Liao, and Rui Mao. R-char: A metacognition-driven framework for role-playing in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26984–27002, 2025.

[229] Zishang Qiu, Xinan Chen, Long Chen, and Ruibin Bai. Mela: A metacognitive LLM-driven architecture for automatic heuristic design. *arXiv preprint arXiv:2507.20541*, 2025.

[230] Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. Recursive introspection: Teaching LLM agents how to self-improve. In *ICML 2024 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*, 2024.

[231] Matthew Renze and Erhan Guven. Self-reflection in large language model agents: Effects on problem-solving performance. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 516–525. IEEE, 2024.

[232] Beatriz A Ribeiro, Helder Coelho, Ana Elisabete Ferreira, and João Branquinho. Metacognition, accountability and legal personhood of ai. In *Multidisciplinary perspectives on artificial intelligence and the law*, pages 169–185. Springer, 2023.

[233] Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. Escalation risks from language models in military and diplomatic decision-making. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, pages 836–898, 2024.

[234] David Rosenthal. Consciousness and metacognition. In *Metarepresentation: Proceedings of the tenth Vancouver cognitive science conference*, pages 265–295. Oxford University Press New York, 2000.

[235] Mahule Roy and Subhas Roy. Physics-informed metacognition: Improving LLMs self-knowledge via physical constraints. In *EurIPS 2025 Workshop on Metacognition in Generative AI*, 2025.

[236] Yinghao Sang. Autocrit: A meta-reasoning framework for self-critique and iterative error correction in LLM chains-of-thought. In *2025 6th International Conference on Machine Learning and Computer Application (ICMLCA)*, pages 1177–1180. IEEE, 2025.

[237] Reuben Sass. Meta-cognitive competence and ai-assisted decision-making: Revisiting the role of explainable AI and uncertainty quantification. *Philosophy & Technology*, 38(3):113, 2025.

[238] M Schmill, Tim Oates, Michael L Anderson, Darsana Josyula, Don Perlis, Shomir Wilson, and Scott Fults. The role of metacognition in robust AI systems. In *Workshop on Metareasoning at the Twenty-Third AAAI Conference on Artificial Intelligence*, 2008.

[239] Florian Scholten, Tobias R Rebholz, and Mandy Hütter. Metacognitive myopia in large language models. *arXiv preprint arXiv:2408.05568*, 2024.

[240] Gregory Schraw. Promoting general metacognitive awareness. *Instructional science*, 26(1): 113–125, 1998.

[241] Gregory Schraw. A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and learning*, 4(1):33–45, 2009.

[242] Bennett L. Schwartz, Tina Sundelin, and Andreas Jemstedt. Meta-metacognition: Processes underlying judgments about metacognition. *New Ideas in Psychology*, 82:101254, 2026. ISSN 0732-118X. doi: https://doi.org/10.1016/j.newideapsych.2026.101254. URL `https://www.sciencedirect.com/science/article/pii/S0732118X2600019X`.

[243] Richard Servajean and Philippe Servajean. Measuring the metacognition of ai. *arXiv preprint arXiv:2603.29693*, 2026.

[244] Ricky J Sethi, Hefei Qiu, Charles Courchaine, and Joshua Iacoboni. Do LLMs dream of electric emotions? towards quantifying metacognition and generalizing the teacher-student model using ensembles of LLMs. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 5223–5227, 2025.

[245] Ruixin Sha, Conghui Sun, Chunliang Yang, Liang Luo, and Xiao Hu. A trade-off between reasoning ability and metacognitive sensitivity in large language models. 01 2026. doi: 10.31234/osf.io/8jq4f_v1.

[246] Paulo Shakarian. Toward artificial metacognition. In *AAAI*, 2026.

[247] Aria Shakoo and Hossein Sameti. Dynamic cognitive orchestration: Eliciting metacognitive planning in large language models, 2025. URL https://openreview.net/forum?id=1GoQe3r0Lv.

[248] Liang Shan, Kaicheng Shen, Wen Wu, Zhenyu Ying, Chaochao Lu, Guangze Ye, and Liang He. Mentor: A metacognition-driven self-evolution framework for uncovering and mitigating implicit risks in LLMs on domain tasks. *arXiv preprint arXiv:2511.07107*, 2025.

[249] Xu Shen, Qi Zhang, Song Wang, Zhen Tan, Xinyu Zhao, Laura Yao, Vaishnav Tadiparthi, Hossein Nourkhiz Mahjoub, Ehsan Moradi Pari, Kwonjoon Lee, et al. Metacognitive self-correction for multi-agent system via prototype-guided next-execution reconstruction. *arXiv preprint arXiv:2510.14319*, 2025.

[250] Marta Siedlecka, Borysław Paulewicz, and Michał Wierzchoń. But i was so sure! metacognitive judgments are less accurate given prospectively than retrospectively. *Frontiers in psychology*, 7:218, 2016.

[251] Significant Gravitas. AutoGPT. URL https://github.com/Significant-Gravitas/AutoGPT.

[252] Andrea Sikora, Leo A. Celi, and Raja-Elie E. Abdulnour. Can ai say "i don't know"? *New England Journal of Medicine*, 394(19):1873–1875, 2026. doi: 10.1056/NEJMp2517624. URL https://www.nejm.org/doi/full/10.1056/NEJMp2517624.

[253] Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. Trust me, i'm wrong: High-certainty hallucinations in LLMs. *arXiv preprint arXiv:2502.12964*, pages arXiv–2502, 2025.

[254] Anjali Singh, Zhitong Guan, and Soo Young Rieh. Enhancing critical thinking in generative AI search with metacognitive prompts. *Proceedings of the Association for Information Science and Technology*, 62(1):672–684, 2025.

[255] Ranganatha Sitaram, Tomas Ros, Luke Stoeckel, Sven Haller, Frank Scharnowski, Jarrod Lewis-Peacock, Nikolaus Weiskopf, Maria Laura Blefari, Mohit Rana, Ethan Oblak, et al. Closed-loop brain training: the science of neurofeedback. *Nature Reviews Neuroscience*, 18 (2):86–100, 2017.

[256] Siyuan Song, Jennifer Hu, and Kyle Mahowald. Language models fail to introspect about their knowledge of language. *arXiv preprint arXiv:2503.07513*, 2025.

[257] Siyuan Song, Harvey Lederman, Jennifer Hu, and Kyle Mahowald. Privileged self-access matters for introspection in ai. *arXiv preprint arXiv:2508.14802*, 2025.

[258] Zhangde Song, Jieyu Lu, Yuanqi Du, Botao Yu, Thomas M. Pruyn, Yue Huang, Kehan Guo, Xiuzhe Luo, Yuanhao Qu, Yi Qu, Yinkai Wang, Haorui Wang, Jeff Guo, Jingru Gan, Parshin Shojaee, Di Luo, Andres M Bran, Gen Li, Qiyuan Zhao, Shao-Xiong Lennon Luo, Yuxuan Zhang, Xiang Zou, Wanru Zhao, Yifan F. Zhang, Wucheng Zhang, Shunan Zheng, Saiyang Zhang, Sartaaj Takrim Khan, Mahyar Rajabi-Kochi, Samantha Paradi-Maropakis, Tony Baltoiu, Fengyu Xie, Tianyang Chen, Kexin Huang, Weiliang Luo, Meijing Fang, Xin Yang, Lixue Cheng, Jiajun He, Soha Hassoun, Xiangliang Zhang, Wei Wang, Chandan K. Reddy, Chao Zhang, Zhiling Zheng, Mengdi Wang, Le Cong, Carla P. Gomes, Chang-Yu Hsieh, Aditya Nandy, Philippe Schwaller, Heather J. Kulik, Haojun Jia, Huan Sun, Seyed Mohamad Moosavi, and Chenru Duan. Evaluating large language models in scientific discovery. *arXiv preprint arXiv:2512.15567*, 2025.

[259] Kaya Stechly, Karthik Valmeekam, Vardhan Palod, Atharva Gundawar, and Subbarao Kambhampati. Beyond semantics: The unreasonable effectiveness of reasonless intermediate tokens. In *First Workshop on Foundations of Reasoning in Language Models*, 2025.

[260] Mark Steyvers and Megan AK Peters. Metacognition and uncertainty communication in humans and large language models. *Current Directions in Psychological Science*, 35:131–139, 2026.

[261] Mark Steyvers, Catarina Belem, and Padhraic Smyth. Generalization of fine-tuned uncertainty communication and metacognition in large language models. *arXiv preprint arXiv:2510.05126*, 2025.

[262] 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*, 7(2):221–231, 2025.

[263] Emily Su, Jessica Bo, Siyi Wu, Ashton Anderson, and Beth Coleman. Do you think GPT will be correct?: Measuring and improving generative AI literacy through metacognition training.

[264] Zexu Sun, Yongcheng Zeng, Erxue Min, Heyang Gao, Bokai Ji, and Xu Chen. Cogrethinker: Hierarchical metacognitive reinforcement learning for LLM reasoning. *arXiv preprint arXiv:2510.15979*, 2025.

[265] Zhongxiang Sun, Qipeng Wang, Weijie Yu, Jingxuan Yang, Haolang Lu, and Jun Xu. Deep search with hierarchical meta-cognitive monitoring inspired by cognitive neuroscience. *arXiv preprint arXiv:2601.23188*, 2026.

[266] Stefan Szeider. What do LLM agents do when left alone? evidence of spontaneous metacognitive patterns. *arXiv preprint arXiv:2509.21224*, 2025.

[267] Tung Duong Ta, Tim Oates, Thien Van Luong, Huan Vu, and Tien Cuong Nguyen. Mdtoc: Metacognitive dynamic tree of concepts for boosting mathematical problem-solving of large language models. *arXiv preprint arXiv:2512.18841*, 2025.

[268] Zhen Tan, Jie Peng, Song Wang, Lijie Hu, Tianlong Chen, and Huan Liu. Tuning-free accountable intervention for LLM deployment – a metacognitive approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25237–25245, Apr. 2025. doi: 10.1609/aaai.v39i24.34710. URL `https://ojs.aaai.org/index.php/AAAI/article/view/34710`.

[269] Zhisheng Tang and Mayank Kejriwal. Humanlike cognitive patterns as emergent phenomena in large language models. *arXiv preprint arXiv:2412.15501*, 2024.

[270] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. The metacognitive demands and opportunities of generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2024.

[271] Michelle Taub, Roger Azevedo, François Bouchet, and Babak Khosravifar. Can the use of cognitive and metacognitive self-regulated learning strategies be predicted by learners' levels of prior knowledge in hypermedia-learning environments? *Computers in Human Behavior*, 39:356–367, 2014.

[272] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL `https://aclanthology.org/2023.emnlp-main.330/`.

[273] Bert Timmermans, Leonhard Schilbach, Antoine Pasquali, and Axel Cleeremans. Higher order thoughts in action: consciousness as an unconscious re-description process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1412–1423, 2012.

[274] Rémi Tison and Tadeusz Zawidzki. Human versus artificial social cognition and metacognition: the normative difference. *AI and Ethics*, 5(5):5255–5271, 2025.

[275] Anjikya Tiwari and Vibhuti Gupta. Self-aware language models: A taxonomy and evaluation of epistemic uncertainty and hallucination mitigation. 2026. doi: 10.21203/rs.3.rs-8589677/v1.

[276] Hayato Tomisu, Junya Ueda, and Tsukasa Yamanaka. The cognitive mirror: A framework for ai-powered metacognition and self-regulated learning. In *Frontiers in Education*, volume 10, page 1697554. Frontiers Media SA, 2025.

[277] Jason Toy, Josh MacAdam, and Phil Tabor. Metacognition is all you need? using introspection in generative agents to improve goal-directed behavior. *arXiv preprint arXiv:2401.10910*, 2024.

[278] Le Tuan Minh Trinh, Le Minh Vu Pham, Thi Minh Anh Pham, and An Duc Nguyen. Metacognitive sensitivity for test-time dynamic model selection. In *First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models*, 2026. URL https://openreview.net/forum?id=wROFSZZXu9.

[279] Kim Uittenhove, Andrew J Ellis, Fabian Mumenthaler, Ioana Gatzka, and Patrick Jermann. Metacognitive reflection in the era of generative ai. 2025.

[280] Raymond Uzwyshyn. Beyond traditional AI IQ metrics: Metacognition and benchmarking for LLMs, AGI, and ASI. 03 2024. doi: 10.13140/RG.2.2.36817.34400.

[281] Rodolfo Valiente and Praveen K. Pilly. Metacognition for unknown situations and environments (muse). *Neural Networks*, 194:108131, 2026. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2025.108131. URL https://www.sciencedirect.com/science/article/pii/S0893608025010111.

[282] Marcel VJ Veenman, Bernadette HAM Van Hout-Wolters, and Peter Afflerbach. Metacognition and learning: Conceptual and methodological considerations. *Metacognition and learning*, 1 (1):3–14, 2006.

[283] Emilio Villa-Cueva, S M Masrur Ahmed, Rendi Chevi, Jan Christian Blaise Cruz, Kareem Elzeky, Fermin Cristobal, Alham Fikri Aji, Skyler Wang, Rada Mihalcea, and Thamar Solorio. MoMentS: A comprehensive multimodal benchmark for theory of mind. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22591–22611, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1230. URL https://aclanthology.org/2025.findings-emnlp.1230/.

[284] Moritz Von Zahn, Lena Liebich, Ekaterina Jussupow, Oliver Hinz, and Kevin Bauer. Knowing (not) to know: Explainable artificial intelligence and human metacognition. *Information Systems Research*, 2025.

[285] Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *International Conference on Machine Learning*, pages 35413–35425. PMLR, 2023.

[286] Ziyu Wan, Yunxiang Li, Xiaoyu Wen, Yan Song, Hanjing Wang, Linyi Yang, Mark Schmidt, Jun Wang, Weinan Zhang, Shuyue Hu, et al. Rema: Learning to meta-think for LLMs with multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 38: 126621–126667, 2026.

[287] Anqi Wang, Zhengyi Li, Lan Luo, Xin Tong, and Pan Hui. Reflexa: Uncovering how LLM-supported reflection scaffolding reshapes creativity in creative coding. *arXiv preprint arXiv:2601.17769*, 2026.

[288] Guoqing Wang, Wen Wu, Guangze Ye, Zhenxiao Cheng, Xi Chen, and Hong Zheng. Decoupling metacognition from cognition: A framework for quantifying metacognitive ability in LLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25353–25361, Apr. 2025. doi: 10.1609/aaai.v39i24.34723. URL https://ojs.aaai.org/index.php/AAAI/article/view/34723.

[289] Hanbin Wang, Jingwei Song, Jinpeng Li, Qi Zhu, Fei Mi, Ganqu Cui, Yasheng Wang, and Lifeng Shang. Teaching large reasoning models effective reflection. *arXiv preprint arXiv:2601.12720*, 2026.

[290] Haoyu Wang, Tao Li, Zhiwei Deng, Dan Roth, and Yang Li. Devil's advocate: Anticipatory reflection for LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 966–978, 2024.

[291] Jason Z Wang. Mirror: A hierarchical benchmark for metacognitive calibration in large language models. *arXiv preprint arXiv:2604.19809*, 2026.

[292] Jing Yi Wang, Nicholas Sukiennik, Tong Li, Weikang Su, Qianyue Hao, Jingbo Xu, Zihan Huang, Fengli Xu, and Yong Li. A survey on human-centric LLMs. *arXiv preprint arXiv:2411.14491*, 2024.

[293] Yuqing Wang and Yun Zhao. Metacognitive prompting improves understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1914–1926, 2024.

[294] Hua Wei, Paulo Shakarian, Christian Lebiere, Bruce Draper, Nikhil Krishnaswamy, and Sergei Nirenburg. Metacognitive ai: Framework and the case for a neurosymbolic approach. In *International Conference on Neural-Symbolic Learning and Reasoning*, pages 60–67. Springer, 2024.

[295] Bingbing Wen, Chenjun Xu, HAN Bin, Robert Wolfe, Lucy Lu Wang, and Bill Howe. From human to model overconfidence: Evaluating confidence dynamics in large language models. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*, 2024.

[296] Pengcheng Wen, Jiaming Ji, Chi-Min Chan, Juntao Dai, Donghai Hong, Yaodong Yang, Sirui Han, and Yike Guo. Thinkpatterns-21k: A systematic study on the impact of thinking patterns in LLMs. *arXiv preprint arXiv:2503.12918*, 2025.

[297] Piotr Winkielman and Jonathan W Schooler. Consciousness, metacognition, and the unconscious. *The Sage handbook of social cognition*, pages 54–74, 2012.

[298] Philip H. Winne. A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences*, 8(4):327–353, 1996. ISSN 1041-6080. doi: https://doi.org/10.1016/S1041-6080(96)90022-9. URL `https://www.sciencedirect.com/science/article/pii/S1041608096900229`.

[299] Abigail C Wright, Emma Palmer-Cooper, Matteo Cella, Nicola McGuire, Marcella Montagnese, Viktor Dlugunovych, Chih-Wei Joshua Liu, Til Wykes, and Corinne Cather. Experiencing hallucinations in daily life: the role of metacognition. *Schizophrenia Research*, 265:74–82, 2024.

[300] George Alexander Wright. A treasure map to metacognition. In *International Conference on Artificial General Intelligence*, pages 326–337. Springer, 2025.

[301] Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. A survey of uncertainty estimation methods on large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21381–21396, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1101. URL `https://aclanthology.org/2025.findings-acl.1101/`.

[302] Yang Xiang, Yixin Ji, Ruotao Xu, Dan Qiao, Zheming Yang, Juntao Li, and Min Zhang. When is thinking enough? early exit via sufficiency assessment for efficient reasoning. *arXiv preprint arXiv:2604.06787*, 2026.

[303] Yingsha Xie, Tiansheng Huang, Enneng Yang, Rui Min, Wenjie Lu, Xiaochun Cao, Naiqiang Tan, and Li Shen. Mitigating safety tax via distribution-grounded refinement in large reasoning models. *arXiv preprint arXiv:2602.02136*, 2026.

[304] 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. URL https://openreview.net/forum?id=gjeQKFxFpZ.

[305] Jiexi Xu and Qianhui Lu. Agentic metacognition: Designing a self-aware low-code agent for failure prediction and human handoff. In *Proceedings of the Computational Methods in Systems and Software*, pages 410–421. Springer, 2025.

[306] Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. To believe or not to believe your LLM. *arXiv preprint arXiv:2406.02543*, 2024.

[307] Noam Steinmetz Yalon, Ariel Goldstein, Liad Mudrik, and Mor Geva. Indications of belief-guided agency and meta-cognitive monitoring in large language models. *arXiv preprint arXiv:2602.02467*, 2026.

[308] Hanqi Yan, Linhai Zhang, Jiazheng Li, Zhenyi Shen, and Yulan He. Position: LLMs need a bayesian meta-reasoning framework for more robust and generalizable reasoning. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

[309] Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. On verbalized confidence scores for LLMs. *arXiv preprint arXiv:2412.14737*, 2024.

[310] Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E Gonzalez, and Bin Cui. Buffer of thoughts: Thought-augmented reasoning with large language models. *Advances in Neural Information Processing Systems*, 37:113519–113544, 2024.

[311] Wei Yang, Defu Cao, Jiacheng Pang, Muyan Weng, and Yan Liu. Adaptive collaboration with humans: Metacognitive policy optimization for multi-agent LLMs with continual learning. *arXiv preprint arXiv:2603.07972*, 2026.

[312] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

[313] Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. Are reasoning models more prone to hallucination? *arXiv preprint arXiv:2505.23646*, 2025.

[314] Jewon Yeom, Jaewon Sok, Seonghyeon Park, Jeongjae Park, and Taesup Kim. Epicar: Knowing what you don't know matters for better reasoning in LLMs. *arXiv preprint arXiv:2601.06786*, 2026.

[315] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuan-Jing Huang. Do large language models know what they don't know? In *Findings of the association for Computational Linguistics: ACL 2023*, pages 8653–8665, 2023.

[316] Gal Yona, Roee Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764, 2024.

[317] Gal Yona, Mor Geva, and Yossi Matias. Hallucinations undermine trust; metacognition is a way forward. *arXiv preprint arXiv:2605.01428*, 2026.

[318] Andria Young and Jane D Fry. Metacognitive awareness and academic achievement in college students. *Journal of the Scholarship of Teaching and Learning*, 8(2):1–10, 2008.

[319] Bhada Yun, Evgenia Taranova, and April Yi Wang. Does my chatbot have an agenda? understanding human and AI agency in human-human-like chatbot interaction. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–32, 2026.

[320] Zhongshen Zeng, Pengguang Chen, Shu Liu, Haiyun Jiang, and Jiaya Jia. Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation. In *International Conference on Learning Representations*, volume 2025, pages 101693–101714, 2025.

[321] Geng Zhang, Yizhou Ying, Sihang Jiang, Jiaqing Liang, Guanglei Yue, Yifei Fu, Hailin Hu, and Yanghua Xiao. From remembering to metacognition: Do existing benchmarks accurately evaluate LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13440–13457, 2025.

[322] Lanxue Zhang, Yuqiang Xie, Fang Fang, Fanglong Dong, Rui Liu, and Yanan Cao. Metagdpo: Alleviating catastrophic forgetting with metacognitive knowledge through group direct preference optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(41): 34755–34763, Mar. 2026. doi: 10.1609/aaai.v40i41.40777. URL `https://ojs.aaai.org/index.php/AAAI/article/view/40777`.

[323] Simpson Zhang, Tennison Liu, and Mihaela van der Schaar. Agents require metacognitive and strategic reasoning to succeed in the coming labor markets. *arXiv preprint arXiv:2505.20120*, 2025.

[324] Weiying Zhang, Yuhua Xu, and Zhixin Sun. Mira: An LLM-driven dual-loop architecture for metacognitive reward design. *Systems*, 13(12):1124, 2025.

[325] Xuanming Zhang, Yuxuan Chen, Samuel Yeh, and Sharon Li. Metamind: Modeling human social thoughts with metacognitive multi-agent systems. *Advances in Neural Information Processing Systems*, 38:26133–26178, 2026.

[326] Yanbo Zhang, Sumeer A. Khan, Adnan Mahmud, Huck Yang, Alexander Lavin, Michael Levin, Jeremy Frey, Jared Dunnmon, James Evans, Alan Bundy, Saso Dzeroski, Jesper Tegner, and Hector Zenil. Advancing the scientific method with large language models: From hypothesis to discovery. *arXiv preprint arXiv:2505.16477*, 2025.

[327] Muyang Zhao, Qi Qi, and Hao Sun. Roi-reasoning: Rational optimization for inference via pre-computation meta-cognition. *arXiv preprint arXiv:2601.03822*, 2026.

[328] Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen. Fact-and-reflection (FaR) improves confidence calibration of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8702–8718, Bangkok, Thailand, August 2024. doi: 10.18653/v1/2024.findings-acl.515.

[329] Longwei Zheng, Anna He, Changyong Qi, Haomin Zhang, and Xiaoqing Gu. Cognitive echo: Enhancing think-aloud protocols with LLM-based simulated students. *British Journal of Educational Technology*, 56(5):2019–2042, 2025.

[330] Yunxuan Zheng, Samuel Recht, and Dobromir Rahnev. Common computations for metacognition and meta-metacognition. *Neuroscience of Consciousness*, 2023(1):niad023, 01 2023. ISSN 2057-2107. doi: 10.1093/nc/niad023. URL `https://doi.org/10.1093/nc/niad023`.

[331] Kaitlyn Zhou, Dan Jurafsky, and Tatsunori B Hashimoto. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, 2023.

[332] Kaiwen Zhou, Wenli Du, and Xufei Tian. Agentic workflows generation based on metacognitive chain-of-thought guided monte carlo tree search. In *2025 7th International Conference on Natural Language Processing (ICNLP)*, pages 302–306. IEEE, 2025.

[333] Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Yi Fang, Sirui Ding, Jiashuo Wang, Kaishuai Xu, et al. Large language models for disease diagnosis: A scoping review. *npj Artificial Intelligence*, 1(1):9, 2025.

[334] Yujia Zhou, Zheng Liu, Jiajie Jin, Jian-Yun Nie, and Zhicheng Dou. Metacognitive retrieval-augmented large language models. In *Proceedings of the ACM Web Conference 2024*, pages 1453–1463, 2024.