Do Large Language Models Possess Human-like Cognitive Abilities: A Systematic Review and Robustness Testing Based on Classical Psychological Tasks

Authors: Zhong Shuhang, Song Jiaqi, Li Jingyao

Date: 2026-05-08T01:02:48+00:00

Abstract

Large Language Models (LLMs) have demonstrated exceptional performance in natural language processing, sparking extensive discussion within the field of psychology regarding whether they possess human-like cognitive abilities. This study systematically reviews 101 empirical papers evaluating LLMs across five classic cognitive domains: creative thinking, memory monitoring and metacognition, reasoning, problem-solving and executive control, and Theory of Mind. Building upon this, the study utilizes the latest GPT-5.4 model to conduct replication tests and minor perturbation tests on 14 representative studies to deeply assess the robustness of their cognitive abilities. The research results indicate that while mainstream LLMs often achieve scores approaching or even exceeding human baselines in standardized, static classic psychological tasks, they exhibit a high degree of framing effects. Their performance undergoes severe degradation when faced with changes in task format, minor semantic perturbations, or complex scenarios requiring dynamic information updates. This study suggests that the cognitive abilities currently demonstrated by large language models lean more toward explicit behavioral simulation based on linguistic representations and have not yet formed a stable, human-like cognitive system with endogenous mechanisms.

Full Text

Preamble

Do Large Language Models Possess Human-like Cognitive Abilities: A Systematic Review and Robustness Testing Based on Classical Psychological Tasks (School of Psychology, Shenzhen University, Shenzhen 518060)

Abstract: The exceptional performance of Large Language Models (LLMs) in natural language processing has sparked extensive debate within the field of psychology regarding whether these models possess human-like cognitive abilities. This study systematically reviews 101 empirical papers evaluating LLMs across five classical cognitive domains: creative thinking, memory monitoring and metacognition, reasoning, problem-solving and executive control, and Theory of Mind (ToM). Building upon this review, the study utilizes the latest GPT-4 model to conduct replication tests and minor perturbation analyses on 14 representative studies to deeply evaluate the robustness of these cognitive abilities. The results indicate that while mainstream LLMs often achieve scores approaching or even exceeding human baselines in standardized, static classical psychological tasks, they exhibit significant framing effects. Their performance declines sharply when faced with changes in task format, minor semantic perturbations, or complex scenarios requiring dynamic information updates. This research suggests that the cognitive abilities currently demonstrated by LLMs are more likely explicit behavioral simulations based on linguistic representations, rather than the formation of a stable, human-like cognitive system with endogenous mechanisms.

1 引言

In recent years, the rapid development of large language models (LLMs) has demonstrated unprecedented capabilities in tasks such as language understanding, text generation, question answering, and reasoning. Consequently, LLMs have become a significant research theme at the intersection of psychology, cognitive science, and artificial intelligence. Language has long served as a core material in psychological research; because LLMs learn and generate language at an unprecedented scale, they may serve both as new objects of psychological inquiry and as novel platforms for testing cognitive theories (Binz & Schulz, 2023; Hagendorff, Dasgupta, et al., 2023;

R. Xu et al., 2024).

According to Marr' s (2010) levels of analysis, complex information-processing systems can be studied simultaneously across multiple levels. Within this framework, the systematic measurement of behavioral performance holds substantial theoretical value for understanding complex intelligent systems (Marr, 2010). Accordingly, psychological research on these models focuses on whether their performance across classic cognitive constructs is stable and interpretable, and whether such performance can provide insights into the operational mechanisms of human cognitive systems. Existing research has already found that LLMs exhibit compelling performance across several psychological tasks. Binz and Schulz (2023) were among the first to use classic experimental tasks from cognitive psychology to examine GPT-3, finding that it performed similarly to, and in some cases better than, humans in decision-making, information searching, and certain reasoning tasks (Binz & Schulz, 2023). Hagendorff et al. (2023) further discovered that larger language models exhibit human-like intuitive bi-

ases during these tasks (Hagendorff, Fabi, et al., 2023). Research in the field of creativity has also shown that GPT-4

demonstrates high levels of fluency, flexibility, and originality on the Torrance Tests of Creative Thinking (Guzik et al., 2023). In the domain of social cognition, research centered on false-belief tasks suggests that newer models may now be able to pass classic Theory of Mind (ToM) tasks to a certain extent (Marchetti et al., 2023). These results indicate that the performance of LLMs in certain classic psychological tasks warrants empirical investigation.

However, inconsistencies in early evidence have led to ongoing controversy regarding whether LLMs truly possess the corresponding cognitive abilities. For instance, through minor perturbation experiments, Ullman (2023) pointed out that the success of LLMs in Theory of Mind tasks may be highly sensitive to surface forms; minor rewording that preserves the core logic can significantly alter the results (Ullman, 2023).

Furthermore, although the literature in this field is growing rapidly, it remains generally fragmented, with a lack of systematic integration across different domains and paradigms. Existing systematic reviews still primarily focus on single domains, and there is a notable absence of cross-domain integrated research that covers multiple cognitive areas using classic psychological paradigms as a primary framework (Du, 2025).

Based on this context, the present study focuses on five domains: Theory of Mind, creative thinking, reasoning, problem-solving and executive control, and memory monitoring and metacognition. It aims to systematically review the literature from the past two years concerning psychological testing of LLMs. Building on this foundation, this study further adopts a meta-analytic approach to re-evaluate previous research through replication verification and minor perturbation testing. We attempt to answer two core questions: first, which cognitive results remain relatively stable across different models and measurement conditions; and second, what patterns emerge regarding the boundaries and instabilities of LLM capabilities, and what insights these patterns provide for psychological theory and future human-computer collaboration.

2.1 总体思路

This study employs a meta-analytical approach to conduct a systematic review following the PRISMA guidelines. The primary objective is to systematically organize the performance of current Large Language Models (LLMs) across classic psychological cognitive paradigms. Based on this synthesis, we aim to identify representative studies suitable for further retesting and micro-perturbation analysis.

Unlike evaluation methodologies centered on model scores or comprehensive benchmark rankings, this research focuses on whether LLMs exhibit measurable capabilities within cognitive constructs that are psychologically well-defined, the-

oretically grounded, and characterized by stable task boundaries. Consequently, the target literature for this study generally follows a trajectory of "cognitive domain—psychological construct—classic psychological paradigm/task—measurement study transferred to LLMs," rather than engineering-focused research that prioritizes "model—benchmark—score" metrics.

2.2 界定研究范围

Before commencing the retrieval process, we defined the scope of our research. This study focuses on the cognitive capabilities exhibited by existing, widely-used large language models (LLMs) during practical tasks. We specifically exclude performance enhancements achieved through task-specific fine-tuning, complex prompting frameworks, external tool integration, or multi-agent collaboration. Our objective is to address the question: "Do the large language models accessible to the general public demonstrate measurable cognitive abilities in their native state?" Consequently, this research prioritizes the direct measurement of official model performance and does not consider engineering-based enhancement solutions as core evidence.

In addition, this study excludes purely theoretical articles and research lacking empirical test data. However, review articles that possess significant academic value and provide comprehensive summaries of existing methodologies are included within the scope of this analysis.

[Figure 1: see original paper]

2 Methodology

2.1 Data Collection and Selection Criteria

To ensure the rigor and representativeness of the literature sample, we established a multi-stage screening process. First, we conducted a systematic search across major academic databases, including IEEE Xplore, Web of Science, and arXiv, using keywords related to machine learning and deep learning applications in the field.

The initial search yielded a broad set of results, which were then subjected to the following inclusion and exclusion criteria: - **Inclusion Criteria:** Peer-reviewed journal articles and conference papers; studies providing clear experimental setups; and research utilizing standardized datasets for performance evaluation. - **Exclusion Criteria:** Non-English publications; short abstracts or posters; and studies where the methodology was insufficiently detailed to allow for reproducibility.

2.2 Mathematical Framework

The core of our comparative analysis relies on a unified mathematical representation of the optimization problems discussed in the literature. For a given

objective function $f(x)$, we consider the general optimization problem:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_j(x) = 0, \quad j = 1, \ldots, p \end{aligned}$$

In the context of deep learning, the objective function $f(x)$ typically represents a loss function $\mathcal{L}(\theta)$, where θ denotes the model parameters. The optimization process aims to find the optimal parameter set θ^* that minimizes the empirical risk over the training data. As noted in [**?**], the gradient descent update rule can be expressed as:

$$\theta_{t+1} = \theta_t - \eta \nabla \mathcal{L}(\theta_t)$$

where η represents the learning rate. We further analyze variations of this approach, such as Adam and RMSProp, which incorporate momentum and adaptive learning rates to improve convergence stability in high-dimensional parameter spaces.

2.3 Performance Metrics

Literature that clarifies theoretical backgrounds and methodologies was retained. Furthermore, this study does not include dynamic video tasks or research focused solely on Visual Language Models (VLMs) as core subjects. The exclusion of these studies is based on the fact that they typically introduce additional factors, such as perceptual interaction and the use of external tools, which significantly increase task heterogeneity. Such complexity makes it difficult to trace research results back to a single psychological construct. In contrast, pure text tasks and a limited number of static image tasks with clear structures and no external tool integration are more aligned with the objectives of this research.

Finally, following the initial round of experimental retrieval, we decided to incorporate complex, large-scale comprehensive benchmarks primarily as peripheral evidence rather than as core evidence for constructing our primary conclusions. This decision stems from the fact that comprehensive benchmarks often involve extensive rewriting and expansion of classical tasks, adaptations of original scoring rules, and the introduction of additional engineering control variables. These modifications tend to weaken the direct correspondence between the benchmarks and the original psychological paradigms. Furthermore, large-scale comprehensive benchmarks frequently aggregate subtasks from multiple distinct constructs into a single total score. While this approach facilitates horizontal comparisons of performance across different models, the resulting constructs are often overly heterogeneous, leading to insufficiently clear interpretative boundaries from a psychological perspective.

2.3 文献检索

This study conducted two rounds of literature retrieval. The first round, an exploratory search conducted between September and October 2024, aimed to map the overall distribution of literature in the field and refine the search strategy accordingly. During this initial phase, the research team had not yet systematically incorporated keywords related to psychological paradigms, nor had they excluded studies focused on fine-tuning or prompt engineering enhancements.

Following a preliminary screening, the team observed that the relevant literature was both voluminous and highly heterogeneous. A significant portion of these studies focused on fine-tuning techniques, prompt engineering, complex multi-agent collaboration, or the use of novel tasks designed by the researchers themselves. Although many of these papers claimed to evaluate capabilities such as reasoning and creative thinking, they often lacked a clear foundation in established psychological constructs. Furthermore, these custom-designed tasks proved difficult to utilize for cross-study comparisons. Based on these findings, the research team shifted the focus of the search terms toward the application of classical psychological paradigms, prioritizing the inclusion of studies that explicitly adapted established tasks associated with specific psychological constructs.

Research has shifted toward Large Language Models (LLMs) to enhance both the interpretability of literature integration and the operability of subsequent replication tests. Based on this philosophy, we conducted a formal second search. This search was organized around five primary domains and their corresponding secondary domains. For each secondary domain under a primary domain, we constructed search terms related to its constructs and classic experimental paradigms. Furthermore, drawing on experience from the pilot search phase, we incorporated relevant exclusion terms to improve the specificity of the search results. Given the rapid, day-to-day evolution of the LLM field, we maintained dynamic updates for new literature alongside the main search; consequently, the search period spanned from April 2025 to March 2026.

We have also applied temporal constraints to the target literature. During the experimental retrieval phase, we initially focused on the early period of 2023, when large-scale models were first released.

The starting point for this study was set following the release of GPT-4. During the formal retrieval phase, the research team primarily narrowed the search timeframe to literature published after January 1,

2024. This decision was made because large language models iterate extremely rapidly, and early conclusions can quickly lose their relevance. However, several foundational studies identified during the initial search were retained due to their enduring significance to the field.

The formal search was primarily conducted across four databases: PsycINFO, Web of Science Core Collection, Scopus, and arXiv. The disciplinary scope was

restricted to psychology, cognitive science, neuroscience, and computer science. From these databases, we initially retrieved 261 papers on Theory of Mind, 365 on creative thinking, 639 on reasoning, 913 on problem-solving and executive control, and 533 on memory monitoring and metacognition. After removing duplicates, the final counts for these categories were 234, 339, 621, 889, and 516 papers, respectively.

2.4 文献筛选

The literature screening process was conducted in two stages following the PRISMA framework. The first stage involved screening titles and abstracts, assisted by ASReview Lab, and was performed independently by two researchers. Prior to the formal screening, the two researchers conducted a consistency check on a sample of 1,200 articles. The resulting Cohen' s kappa coefficient was 0.866, indicating a high level of agreement in their inclusion and exclusion judgments.

The objective of this stage was to identify studies directly related to classical psychological cognitive paradigms that warranted further full-text review from the large-scale search results, based on established inclusion and exclusion criteria. Following the first round of screening, the number of documents retained in each primary domain was as follows: 22 for Theory of Mind, 32 for Creative Thinking, 41 for Reasoning, 27 for Problem Solving and Executive Control, and 15 for Memory and Metacognition.

The second stage of screening was conducted at the full-text level. The research team read the full text of the documents retained from the first round to further determine whether they met the core criteria of this study. These criteria included: whether the measurement targets were clearly defined psychological constructs or paradigms, whether the task design had a sufficiently clear orientation toward specific abilities, and whether the results could support subsequent systematic review and integration.

Furthermore, the team assessed whether the materials, prompts, and scoring rules were accessible. After the second round of full-text screening, a total of 101 documents were ultimately included for the subsequent stages of coding, classification, and comprehensive analysis.

2.5 数据提取与编码

For studies that passed the full-text screening, this research established a comprehensive data extraction framework to create a long-form database, systematically coding each literature source following meta-analytical principles. The extracted information includes the authors, publication year, primary and secondary research fields, core research content, names and number of test tasks, task conditions, the specific abilities measured and their underlying psychological paradigms, descriptions of task formats, whether benchmarks were utilized, the testing language, the models involved, item types, scoring methods and met-

rics, the total number of test items, the number of repeated trials, quantitative test results, and the textual interpretation of these findings.

and conclusions. This systematic data collection provides a rigorous foundation for subsequent cross-study comparisons and future replicability verification.

2.6 遴选重复性检验与微小扰动测试文献

Building upon the systematic review, this study further selected two categories of candidate literature for subsequent replicability testing and minor perturbation testing. The replicability testing primarily focuses on studies concluding that models lack a specific capability, possess poor or unstable capabilities, or perform significantly worse than humans. The objective is to examine whether these conclusions remain valid when applied to the latest mainstream models. In this selection process, we prioritized studies featuring simple and direct tasks with clear capability indicators, as well as those whose original test materials, prompts, and scoring rubrics demonstrate high reproducibility.

The minor perturbation testing is primarily directed toward studies claiming that models have already acquired a certain capability and perform well or relatively stably on classical paradigms. These tests aim to maintain the original logic of the questions while introducing slight perturbations—such as replacing character names, paraphrasing descriptions, adjusting sequences, adding or removing irrelevant information, or changing option formats—to evaluate the model' s robustness (Ullman, 2023). For this category, we favored tasks that do not rely on complex prompting frameworks and are well-suited for minor perturbations.

Based on these criteria, we ultimately selected 10 articles for replicability testing and 4 articles for minor perturbation testing.

3.1 研究范围界定

In psychology, originality and utility are regarded as the minimum criteria for creative output. That is, a product must both deviate from convention and be meaningful and useful within a specific task, context, or goal (Runco & Jaeger, 2012).

Based on this conceptual framework, several psychological paradigms have been developed to assess creative thinking. Divergent thinking tasks emphasize the generation of multiple, diverse, and relatively novel responses to open-ended prompts (Guilford, 1967; Runco & Jaeger, 2012). Conversely, convergent thinking tasks focus on the integrative ability to appropriately link multiple distant concepts (Mednick, 1962). Beyond these standard psychological paradigms, practical research methods such as creative writing, idea generation, and literary text evaluation also exist. Although these methods possess high ecological validity, their outcomes are influenced by a mixture of linguistic expression, domain knowledge, aesthetic judgment, and rater preference. Consequently, they are

included only as peripheral specialized sections rather than as core evidence for whether Large Language Models (LLMs) possess psychological creative thinking.

The primary paradigms for measuring divergent thinking in psychology include the Alternative Uses Task (AUT) and the verbal components of the Torrance Tests of Creative Thinking (TTCT). The AUT requires participants to propose as many unconventional uses as possible for common objects; traditional scoring typically includes fluency, flexibility, originality, and elaboration. The verbal tasks within the TTCT encompass open-ended item types such as the Consequences Task (CT) and the Divergent Association Task (DAT) (Guilford, 1967; Torrance, 1974).

The primary paradigm for measuring convergent thinking in psychology is the Remote Associates Test (RAT), found within the TTCT verbal tasks, along with related associative integration tasks. The RAT requires participants to identify a common word that can simultaneously link three seemingly unrelated cue words. It is therefore widely regarded as a critical tool for measuring the ability to integrate distant concepts and achieve convergent creativity (Mednick, 1962). Building upon this foundation, subsequent research has developed various insight problems and word association tasks to examine an individual' s ability to move beyond surface representations and achieve integration within a distant semantic space (Bowden & Beeman, 2003).

3.2 综述与方法学研究

Before proceeding to the empirical study, we first examine several recent reviews and methodological evaluations in the field. A systematic review and meta-analysis by Holzner et al. (2025) reveals an uneven distribution of evidence in current research on generative AI and creativity. While work closely aligned with classical psychological traditions focuses primarily on divergent thinking tasks such as the Alternative Uses Task (AUT), Consequences Test (CT), and Divergent Association Task (DAT), a significant body of subsequent research has shifted toward creative writing and business idea generation. Consequently, the conceptualization of creative thinking across different studies has become conflated with multi-dimensional evaluative factors (Holzner et al., 2025). Similarly,

R. Li et al. (2025), Lu et al. (2026), and Anca et al. (2025) further elucidate the boundaries of current testing from the perspectives of automated scoring methods, evaluation metrics, and test design, respectively.

R. Li et al. (2025) demonstrate that the validity of using Large Language Models (LLMs) for automated creativity assessment depends largely on the reference standards and scoring rubrics provided by researchers. Furthermore, Anca et al. (2025) and Lu et al. (2026) caution that many existing creativity assessments have drifted away from classical psychological paradigms, evolving instead into measures of textual representation—such as linguistic novelty or structural com-

plexity—rather than the creative process itself (Anca et al., 2025; Lu et al., 2026).

Collectively, these studies point toward a critical trajectory: the operational definition of creative thinking is undergoing a shift from measuring cognitive processes to evaluating textual outputs. While this transition imbues assessment tasks with greater real-world relevance, it also risks simplifying the psychological core of creativity into mere textual pattern matching. When the validity of automated scoring is highly constrained by the a priori settings of scoring rules, we are essentially evaluating the degree to which a model aligns with specific rhetorical logics. This current state of measurement—targeting representation rather than essence—underscores the urgent need to re-anchor classical psychological paradigms within the context of artificial intelligence.

3.3 发散思维

Direct Transfer Studies of Classical Divergent Thinking Paradigms

In early research, Gilhooly (2024) conducted a comprehensive review of studies centered on the Alternate Uses Task (AUT). The findings indicated that large language models, such as GPT-3, GPT-4, and ChatGPT, often achieve performance levels in these tasks that are comparable to, or even exceed, the average human performance.

Notably, these models stand out in metrics such as originality, utility, and output fluency. Some findings further indicate that LLM outputs may exhibit behavioral patterns similar to those observed in human Alternative Uses Tasks (AUT); for instance, responses generated later in a sequence tend to be more novel, and a certain trade-off exists between originality and utility [**?**, **?**]. However, Gilhooly [**?**] also emphasizes that such similar patterns do not directly imply that the models undergo the same processes of episodic memory retrieval, object property processing, strategy switching, and executive control as humans.

Hubert et al. (2024) compared the performance of GPT-4 against 151 human participants across three types of tasks: the Alternative Uses Task (AUT), the Consequences Task (CT), and the Divergent Association Task (DAT). Their results demonstrated that GPT-4 achieved higher scores in originality and elaboration for the AUT and CT, and exhibited greater semantic distance in the DAT. Notably, this study quantified the specific advantages of the model in divergent thinking tasks through the measurement of semantic distance.

The results indicate that content generated by the model exhibits higher heterogeneity, novelty, and a greater degree of semantic expansion within the scoring system [**?**]. Arora et al. [**?**] compared the performance of ChatGPT-4o, DeepSeek-V3, and Gemini 2.0 against students on the Alternative Uses Task (AUT) and the Remote Associates Test (RAT). Their findings revealed that all three Generative AI (GenAI) models surpassed the student samples in terms of

both average idea originality and top-tier idea originality on the AUT [**?**]. Similarly, Latif et al. [**?**] directly measured the performance of OpenAI o1-preview on AUT and RAT tasks within the framework of Higher-Order Thinking Skills (HOTS) in education. By comparing the model' s output to human subject performance data extracted from previous studies, the researchers found that the model achieved higher originality scores on the AUT than university students and also exceeded human benchmarks on the RAT [**?**].

However, the limitations of Large Language Models (LLMs) in creative thinking remain significant. Koivisto and Grassini (2025) reviewed the performance of ChatGPT-3.5, ChatGPT-4, and CopyAi compared to humans in the Alternative Uses Task (AUT). Their findings indicated that while AI outperformed humans in terms of average scores and certain maximum value metrics, the highest scores for peak originality were still achieved by humans [**?**, **?**].

Haase and Hanel (2025) also found that the average performance of several Large Language Models (LLMs) on the Alternative Uses Task (AUT) exceeded the human average. However, they further noted that out of 1,061 model responses, only three reached the top 10% of human creativity levels. Furthermore, they observed significant variance in performance across repeated sampling of the same model [**?**, **?**].

Research utilizing the Divergent Association Task (DAT) has further highlighted these performance disparities. Using large-scale human DAT data as a benchmark, Bellemare-Pepin et al. (2024) found that GPT-4' s average DAT score was significantly higher than the overall human average, while Gemini Pro performed near the human mean. However, when the reference group was shifted to highly creative individuals, the models remained below the performance levels of the top 50%, top 25%, and top 10% of human participants [**?**, **?**]. This pattern is consistently observed across various adapted paradigms discussed later in this paper. For instance, results from the PACE task conducted by Qiu and Hu (2025) also demonstrate that while high-level Large Language Models (LLMs) achieve associative distances comparable to those of average humans, professional human experts remain significantly superior. Furthermore, the outputs generated by these models tend to exhibit a higher degree of homogeneity compared to their human counterparts.

Runco et al. (2025) revealed similar limitations by examining the scoring mechanisms of these classical paradigms. They quantified the responses of Bard, GPT-3.5, and GPT-4.0 across various divergent thinking tasks using two specific metrics:

The first is Idea Density, which refers to the number of distinct, identifiable ideas contained within a single response; the second is···

The concept of Semantic Distance is primarily used to evaluate the gap between a model' s innovative responses and conventional answers. Results indicate that prompting a model to be "as original as possible" significantly increases the "idea density" of its output—a phenomenon that appears consistently across

various models. However, metrics for semantic distance are not always stable. In other words, an increase in the quantity of a model' s responses does not necessarily mean that the number of truly independent, creative ideas increases proportionally [**?**].

We might speculate that the high scores achieved by models in certain divergent thinking tasks likely stem from their ability to produce denser, more exhaustive enumerations. This makes them more likely to be recognized by scoring systems as being "idea-rich," without necessarily implying a stronger breakthrough in genuine creativity. Collectively, these studies suggest that when provided with open-ended prompts and explicit scoring criteria, mainstream Large Language Models (LLMs) can perform well on classic developmental assessments.

In divergent thinking tasks, these systems can generate a sufficient quantity of responses that appear novel on the surface and are frequently rated as creative by scoring systems. Furthermore, they consistently achieve above-average performance when compared to human benchmarks. However, the interpretation of these results requires caution. First, tasks such as the Alternate Uses Task (AUT) and the Torrance Tests of Creative Thinking (TTCT) are widely circulated, posing a significant risk of training data contamination. Second, common scoring dimensions—specifically fluency and elaboration—inherently favor systems capable of continuous text generation. We also observe that many studies in this field rely on automated machine scoring or self-evaluation (LLM-as-a-Judge); as discussed in subsequent sections, there are substantial concerns regarding the validity and robustness of these scoring methodologies.

Furthermore, through comparative analysis, it becomes evident that breakthroughs and high-level originality in model responses remain rare. Consequently, while current mainstream Large Language Models (LLMs) have demonstrated strong performance in classical divergent association tasks, we contend that this success fundamentally relies on their statistical advantage in exhaustively traversing permutations within a semantic space. This mechanism remains distinct from the genuine originality characteristic of high-order human creativity.

(2) Research Adapting and Expanding Classical Divergent Thinking Paradigms

Dinu and Florescu (2024) developed an automated scoring measurement framework based on classical paradigms such as the Alternative Uses Task (AUT), Divergent Association Task (DAT), and Consequences Task. Within this framework, they conducted a comparative study between 10 Large Language Models (LLMs) and 10 human participants. The models tested in their default states included Llama-3-70B, Mixtral-8x7B, Cohere command-r-plus, Yichat-34B, Falcon, Copilot (Balanced), Gemini (free), Jais-30B, You.com (Smart mode), and Character AI.

The results indicated that human participants scored higher than the LLMs only

on the Divergent Association Task. In the other four categories of tasks, the average performance of the models exceeded that of the humans [**?**, **?**]. However, several methodological adjustments—such as the researchers creating new test items, imposing word count limits, manually correcting model formatting, and employing reverse scoring for certain tasks—suggest that the validity of this measurement framework may deviate significantly from that of the original classical paradigms.

Zhao et al. (2025) adapted seven types of tasks based on the Torrance Tests of Creative Thinking (TTCT) verbal tasks. Using GPT-4 as an evaluator, they assessed the performance of several mainstream Large Language Models (LLMs) across four classic dimensions: fluency, flexibility, originality, and elaboration.

The evaluated models include GPT-3.5-turbo-0613, LLaMA-2-13b-chat, LLaMA-2-70b-chat, Vicuna-7b-v1.5, Vicuna-13b-v1.5, and Qwen-7b-chat. Significant performance variations were observed across these different models, with GPT-3.5 generally achieving the highest overall creativity scores. Among the four evaluated dimensions, the models scored highest in elaboration and lowest in originality [**?**]. This finding is consistent with previously mentioned conclusions, suggesting that Large Language Models (LLMs) still face limitations when tasked with generating truly rare or unconventional ideas that deviate significantly from established norms.

The CreativityPrism framework, proposed by Hou et al. (2026), integrates classical divergent thinking tasks with creative writing and logical problem-solving into a unified evaluation system. This framework was used to assess 17 Large Language Models (LLMs), including Qwen2.5-72B, GPT-4.1, Gemini 2.0-Flash, Claude 3, DeepSeek-R1, and DeepSeek-V3. Furthermore, the authors developed their assessment criteria based on established classical creativity evaluation frameworks.

Creativity can be decomposed into three distinct evaluative dimensions: Quality, Novelty, and Diversity [**?**, **?**]. A primary contribution of this research is the discovery that multiple metrics obtained by Large Language Models (LLMs) within the same task tend to exhibit high correlations, particularly in Torrance Tests of Creative Thinking (TTCT) type tasks. However, the generalization of these models across different domains and dimensions remains unstable. The authors specifically emphasize that the correlation between the Novelty dimension and other dimensions is weak and may even be negative [**?**, **?**]. This suggests that a model' s superior performance under one specific definition of novelty does not necessarily imply that it will perform well on other novelty-related tasks.

Overall, research in this phase has primarily focused on expanding task types, reorganizing scoring dimensions, and strengthening automated evaluation within the classical divergent thinking framework. While these approaches are generally more effective at differentiating between models, they often introduce various cross-domain capabilities, thereby confounding measurement validity. Never-

theless, these comprehensive adapted tests have yielded conclusions consistent with previous research: Large Language Models (LLMs) demonstrate strong performance in elaboration and certain novel automated metrics, yet they remain limited in terms of genuine, stable, and transferable originality.

3. Practical Tasks Based on the Classical Divergent Thinking Paradigm Framework

Divergent thinking is a core component of creativity, typically characterized by the ability to generate multiple unique solutions to an open-ended problem. To evaluate and cultivate this ability within an academic or computational context, practical tasks are designed based on the classical psychometric paradigms of fluency, flexibility, and originality.

These tasks often utilize the "Alternative Uses Task" (AUT) or "Associative Thinking" models. In a practical application, participants (or models) are presented with a common object—such as a brick or a paperclip—and tasked with generating as many non-traditional uses as possible. This process requires breaking functional fixedness and exploring distant semantic spaces.

The framework for these practical tasks generally follows a three-stage process:

1. **Problem Definition and Stimulus Presentation**: A target stimulus is provided within a specific constraint environment. The goal is to trigger a broad search across the individual' s or system' s knowledge base.
2. **Idea Generation and Expansion**: This stage emphasizes the quantity and variety of responses. In machine learning contexts, this may involve adjusting temperature parameters or utilizing stochastic sampling to ensure the output covers diverse conceptual domains.
3. **Evaluation and Scoring**: Responses are quantified based on established metrics:
    - **Fluency**: The total number of relevant ideas generated.
    - **Flexibility**: The number of different categories or conceptual shifts represented in the responses.
    - **Originality**: The statistical rarity or uniqueness of the ideas compared to a normative database.

By implementing these tasks, researchers can systematically measure the capacity for divergent search and identify the cognitive or algorithmic mechanisms that facilitate creative breakthroughs. This paradigm serves as a foundational benchmark for assessing both human cognitive potential and the generative capabilities of artificial intelligence systems.

In this section, researchers have developed various practice-oriented tasks based on classical theories of divergent thinking. This body of research exhibits an increase in qualitative analysis and a corresponding decrease in quantitative

metrics, allowing for a more multi-dimensional and comprehensive perspective on Large Language Models (LLMs).

Vinchon et al. (2024) adapted the psychological narrative tasks from the Evaluation of Potential Creativity (EPoC) framework. Originally proposed by Barbot et al. (2016) for use with children and adolescents, this framework includes both divergent tasks—which require the generation of multiple story beginnings or endings—and integrative convergent tasks, which require synthesizing a complete story around a given title or specific characters. Upon testing GPT-3.5 and GPT-4, the researchers observed that some stories generated by the models clearly reorganized elements from existing literary works, such as *Alice in Wonderland*, *The Chronicles of Narnia*, and *The Silver Key*. Furthermore, character naming exhibited high levels of repetition, with certain names appearing frequently across multiple independent generation cycles [**?**, **?**].

To some extent, this reflects the fact that model outputs still tend to exhibit fixed templates even after multiple iterations. Furthermore, researchers tasked ChatGPT with···

When acting as a rater to evaluate the creativity of stories, results indicate that model scores show almost no correlation with human ratings and exhibit poor internal consistency [**?**, **?**]. This suggests that, at least for the evaluation of longer and more complex narrative creative products, models cannot reliably replace human judgment of creativity.

Chakrabarty et al. (2024) adapted the four-dimensional creative structure of the Torrance Tests of Creative Thinking (TTCT) into a specialized expert scoring framework for short stories, known as TTCW, to compare the creative performance of professional human authors against Large Language Models (LLMs). After testing GPT-3.5, GPT-4, and Claude 1.3, the researchers found through expert human evaluation that professional authors achieved significantly higher overall pass rates than the models. Furthermore, the stories generated by the models were criticized for being formulaic, possessing thin narratives, and lacking emotional tension (Chakrabarty et al., 2024). It is evident that the high scores achieved by models in classical divergent thinking tests have not successfully transferred to more complex creative tasks.

···on high-quality literary works recognized by experts. Sun et al. (2025) utilized 13 tasks covering three domains—divergent thinking, problem-solving, and creative writing—and reached similar conclusions: while the best responses from Large Language Models (LLMs) in divergent thinking and certain problem-solving tasks can slightly exceed the human average, their overall performance in creative writing remains relatively weak. Furthermore, the textual diversity of the answers generated by these models is significantly lower than that of humans (Sun et al., 2025).

This pattern is equally prevalent across creative tasks in two specific domains. Drawing upon the Alternative Uses Task (AUT) framework, Ruan et al. (2026) developed a scientific idea generation task to investigate whether Large Lan-

guage Models (LLMs) can demonstrate idea generation capabilities related to divergent thinking when provided only with a scientific keyword and no additional context. After comparing the performance of 41 models across 22 scientific fields, the researchers identified distinct domain-specificity among the models. Furthermore, they observed a clear trade-off between originality and practicality in the models' outputs (Ruan et al., 2026).

Tourajmehr et al. (2025) found that in the generation of short Persian literary sentences, the six participating models exhibited varying strengths and weaknesses across the four dimensions of the Torrance Tests of Creative Thinking (TTCT). Overall, these models have not yet reached human levels of performance regarding diversity, literary nuance, and cultural understanding. Furthermore, even high-scoring models tend to repeatedly rely on similar literary imagery and rhetorical templates (Tourajmehr et al., 2025).

Another study utilized a classic paradigm of creative thinking to observe that Large Language Models (LLMs) struggle to break mental sets in open-ended tasks requiring divergent thinking, exhibiting significant functional fixedness. Although these concepts originate from the field of problem-solving, they are introduced here because the authors employed a traditional creative thinking framework.

Desdevises (2025) directly utilized the classic "Egg Task," requiring GPT-4o to propose as many original solutions as possible to prevent an egg from breaking when dropped from a height of 10 meters [**?**]. The researchers not only quantified the number and diversity of the model' s ideas but also categorized the responses into "Fixation Ideas," which fall within dominant conventional pathways, and "Expansion Ideas," which break away from traditional approaches. The results demonstrated that while GPT-4o is capable of generating a large volume of ideas, the vast majority remain concentrated within fixed pathways, with approximately 80.2% classified as fixation ideas. At the same time, the model···

Self-assessments of performance also failed to reliably distinguish between fixed ideas and expansive ideas [**?**, **?**]. This further demonstrates that while the model is prolific, it has not effectively overcome functional fixedness to achieve true originality.

Overall, Large Language Models (LLMs) clearly lose their previous advantages when performing more practical tasks. We can conclude that high scores in classic divergent thinking tasks primarily indicate the model' s ability to generate a large volume of candidate ideas. However, once a task further requires synthesis, breaking cognitive sets, or deep domain-specific expertise, the limitations of these models become significantly more apparent.

3.4 聚合性思维

1. Direct Transfer Research on the Classical Paradigm of Convergent Thinking

Convergent thinking is a core component of human creativity, typically defined as the ability to integrate diverse information to find a single, optimal solution to a well-defined problem. In the field of cognitive psychology, the Remote Associates Test (RAT) serves as the most widely utilized classical paradigm for measuring convergent thinking. Recent research has increasingly focused on the direct transfer of this paradigm into the domain of artificial intelligence, particularly through the lens of machine learning and deep learning.

1.1 Computational Modeling of the Remote Associates Test

The direct transfer of convergent thinking tasks to computational models involves representing linguistic associations within a high-dimensional vector space. Traditional approaches relied on semantic networks and association strength databases; however, modern research leverages large-scale language models (LLMs) to simulate the human "search and retrieval" process. These studies aim to determine whether the associative mechanisms found in human cognition—such as spreading activation—can be effectively replicated through neural network architectures. By applying the RAT paradigm to AI, researchers can quantitatively assess the gap between machine pattern matching and human-like creative synthesis.

1.2 Evaluation Metrics and Performance Benchmarks

When migrating the classical paradigm of convergent thinking to the digital realm, establishing robust evaluation metrics is essential. Beyond simple accuracy (the ability to identify the correct target word), researchers have introduced metrics such as response latency, semantic distance between distractors, and the density of the association space. illustrates the performance comparison between various deep learning models and human baseline groups across standardized RAT datasets. These comparative studies reveal that while machines excel at broad associative retrieval, they often struggle with the "Aha!" moment —the sudden cognitive restructuring characteristic of human insight.

1.3 Constraints and Cognitive Validity

A critical challenge in the direct transfer of convergent thinking paradigms is ensuring cognitive validity. While a machine may solve a convergent thinking task through brute-force search or statistical probability, this does not necessarily imply a functional equivalence to human creative processes. Current research explores how to constrain model parameters to more closely mimic human cognitive limitations, such as working memory capacity and inhibitory control. By incorporating these biological constraints, researchers aim to develop more au-

thentic models of convergent thinking that do not merely solve problems but do so using human-like heuristic strategies. [Figure 1: see original paper] provides a schematic representation of the integrated framework for this transfer process, highlighting the interaction between semantic encoding and the selection of optimal solutions.

In this subfield, there is a notable scarcity of research that directly transfers classical paradigms. As previously mentioned, both Arora et al. (2025) and Latif et al. (2025) employed both the Alternative Uses Task (AUT) and the Remote Associates Test (RAT) to compare mainstream models with human samples. Their results indicate that model performance on RAT tasks has already approached a ceiling. However, the researchers themselves acknowledge that high scores on the RAT can no longer be reliably interpreted as a genuine capacity for remote associative integration. This is due to the high probability of training data contamination resulting from public test bank leaks (Arora et al., 2025). Furthermore, the current dearth of direct classical evidence constitutes an important conclusion in itself: the direct migration of traditional Remote Associates Tests to Large Language Models (LLMs) appears to have lost its scientific utility.

(2) Research Adapting the Classical Paradigm of Convergent Thinking

The core of the research on convergent thinking consists of studies that retain the foundational elements of Mednick' s Remote Associates Theory [**?**], yet deviate significantly from classical problem formats.

Drawing on the free association chain method proposed by Gray et al. (2019), Qiu and Hu (2025) introduced Parallel Association Chain Evaluation (PACE). This framework requires models to generate parallel association chains centered around a specific seed word, after which semantic distance is employed to calculate the associative distance. The researchers found that PACE exhibits a moderate to strong correlation with benchmarks related to creative writing. While the associative distance of high-performing Large Language Models (LLMs) approaches that of the general population, it remains lower than that of professional creative groups. Furthermore, the associations generated by these models tend to be more concrete and exhibit higher levels of homogeneity (Qiu & Hu, 2025).

The CREATE benchmark, introduced by Wadhwa et al. (2026), shifts the focus of testing toward pathfinding by drawing inspiration from the classic Remote Associates Test (Bowden & Beeman, 2003). This benchmark requires models to connect two seemingly unrelated objects or concepts through paths that are as specific and informative as possible. Furthermore, models are expected to maintain diversity and generate a high volume of valid candidate paths. The results indicate that, while frontier models generally outperform others in generating high-quality and rich associative paths, they still struggle to produce truly rare and sufficiently unique connections (Wadhwa et al., 2026).

2026). We argue that these types of exhaustive path-search tasks primarily mea-

sure the statistical advantages of Large Language Models (LLMs) in performing exhaustive permutations within a semantic space.

The LoT benchmark proposed by Huang et al. (2025) is particularly distinctive. This study utilizes *Oogiri*, a traditional Japanese comedy game, as a measurement platform. The task requires models to generate responses based on a given image, text, or a combination of both. To succeed, the model must deviate from conventional reasoning paths to produce an answer that is unexpected yet logical, while simultaneously achieving a humorous effect.

The findings indicate that mainstream models exhibit weak performance in distant creative association and struggle to identify unconventional, creative humor. Even leading models require more than ten rounds of prompting and iterative attempts before their performance approaches the human average [**?**].

Huang et al., 2025). Naeini et al. (2023) adopted a similar design approach, utilizing television game shows as a framework.

The "Connecting Wall" task from the television program *Only Connect* presents a unique challenge for artificial intelligence: models must identify which items from a given set belong to the same group and subsequently articulate the common connection between them. A critical feature of this task is the deliberate inclusion of "red herrings"—items designed to trigger false associations and lead the solver astray. Researchers have explicitly noted that this task borrows from the Remote Associates Test (RAT) paradigm.

Findings indicate that Large Language Models (LLMs) are particularly susceptible to these misleading cues, often falling victim to mental sets [**?**, **?**]. In these practical adaptations of associative tasks, the limitations of LLMs are twofold: they demonstrate a diminished capacity for the "leap" of associative thinking required for success, and they struggle to overcome the effects of functional fixedness and established mental sets.

3.5 创造性思维领域总体评价

By examining these two subfields together, we can derive a relatively comprehensive conclusion. Regarding divergent thinking, substantial evidence of task performance has already been accumulated; many mainstream Large Language Models (LLMs) are not inferior to humans in average-level comparisons and even frequently outperform the average human. However, truly elite, high-end creative performance remains rare. In contrast, direct evidence for convergent thinking is significantly more scarce. Research in this area has primarily focused on highly adapted and engineered tasks, which suggest that LLMs possess a certain capacity for remote association and the linking of distant concepts. Nevertheless, their ability to produce truly high-quality, integrative associations remains limited.

4.1 研究范围界定

This section focuses on two primary research areas: source monitoring and memory monitoring, as well as metacognition. We have intentionally excluded all Large Language Model (LLM) literature broadly labeled with "memory" because our initial pilot search revealed that the majority of existing studies address issues such as pre-trained knowledge, parameter storage, external memory modules, or retrieval-augmented generation. These topics do not align with the psychological definition of memory capacity [**?**, **?**].

Furthermore, while associative or relational memory is occasionally mentioned, there is currently a lack of stable research that strictly corresponds to psychological paradigms. Consequently, these topics do not yet constitute a sufficient basis for an independent inclusion category [**?**, **?**].

Following two rounds of systematic searching and screening, we found no studies that explicitly utilize source monitoring, source memory, or reality monitoring as a theoretical framework, nor any that employ corresponding psychological paradigms to evaluate the ability of LLMs to distinguish between information sources. We observed that while existing literature frequently discusses hallucinations, citation errors, and context confusion, these phenomena are rarely situated within the psychological frameworks of source or reality monitoring. Furthermore, few studies utilize these paradigms to test whether a model can distinguish whether information originates from external input, previous dialogue, or internal model generation [**?**, **?**].

In contrast, a small number of studies concerning working memory appeared in our search results, utilizing relatively clear psychological task formats.

Given that working memory is a core component of the memory system, we have included it as a supplementary subcategory in this section. This inclusion also suggests that while a theory-driven, top-down search strategy helps maintain conceptual boundaries, it may simultaneously reduce the detection rate of closely related concepts.

4.2 元记忆/元认知监控

In psychology, metacognition is typically defined as the monitoring of one' s own knowledge capacity, uncertainty, and response accuracy, as well as the regulation of subsequent behavior based on this monitoring (Butlin, 2026; Steyvers & Peters, 2025). Within psychological research, confidence is generally regarded as the critical variable linking first-order task performance with second-order self-monitoring. Specifically, individuals must decide whether to proceed with a response, abandon it, or seek a more reliable alternative based on their level of confidence in their answer (Fleming, 2024; Kepecs et al., 2008). Consequently, confidence is considered the primary psychological paradigm for studying metacognition. Based on this framework, we have categorized the included literature into four main research directions for discussion.

Literature Review of Previous Research

Among the ten documents included in this study, three review articles provide a foundational research framework for the field. Steyvers and Peters (2025) synthesized the primary methodologies used to adapt human metacognitive research paradigms to Large Language Models (LLMs). The authors distinguish between two core evaluative metrics: metacognitive sensitivity (monitoring sensitivity), which refers to whether the confidence levels reported by a model can effectively discriminate between correct and incorrect responses; and metacognitive calibration, which assesses whether the model' s reported degree of certainty aligns with its overall objective accuracy \cite{Fleming, 2024; Lee et al., 2025;

Z. Li & Steyvers, 2025}.

Furthermore, the authors summarize the prevailing measurement pathways into two categories. The first is implicit measurement, which infers model confidence from internal states such as token likelihood, probability distributions over options, p(true), or consistency across multiple samples. The second is explicit measurement, which involves directly prompting the model to report its certainty regarding a specific question using numerical values or natural language. The authors emphasize that these two measurement approaches may correspond to different levels of cognitive processing and should not be conflated [**?**, **?**].

Researchers have discovered that internal uncertainty signals within Large Language Models (LLMs) can, to some extent, predict which questions are more likely to be answered correctly and which are prone to error [**?**, **?**]. However, the explicit confidence expressions provided by these models are often unstable, and overconfidence remains a prevalent issue [**?**, **?**]. Notably, implicit indicators typically align more closely with actual accuracy than explicit self-reports [**?**, **?**].

The literature also explores the similarities and differences between humans and LLMs in this domain. A key similarity is that both entities can exhibit overconfidence and utilize linguistic markers such as "likely" or "probably" to communicate uncertainty; in some cases, LLM judgments even approximate the average judgment of human groups [**?**, **?**]. Nevertheless, these surface-level similarities suggest only that LLMs exhibit metacognitive-like behaviors in specific tasks. The authors argue that improvements in a single performance metric should not be conflated with a genuine advancement in underlying metacognitive capacity [**?**, **?**].

Butlin (2026) primarily evaluates whether current Large Language Models (LLMs) can form higher-order representations of their own internal representational states. In addressing this question, Butlin (2026) reviews three main categories of existing research. Among those related to metacognitive measurement paradigms are confidence-based studies and self-prediction studies. The latter involves requiring a model to predict how it will respond under specific prompting conditions and comparing this self-predictive ability to its ability to predict the responses of other models [**?**]. While the author

observes phenomena compatible with higher-order representation in both domains, he argues that these results are insufficient to prove that LLMs have formed such representations. This is because most existing research has yet to distinguish whether a model is truly representing its current internal cognitive state or merely utilizing first-order cues associated with task success to produce behavior that mimics self-monitoring [**?**].

To illustrate this point, the author draws on a classic controversy in animal metacognition research: in animal studies, even if a subject is more willing to bet on high-accuracy trials and remains more conservative on low-accuracy trials, it may simply have learned to form conditioned reflexes based on external stimuli rather than monitoring its own knowledge state [**?**, **?**, **?**]. Butlin (2026) contends that current LLM research faces the same type of conceptual validity challenge.

Zhang et al. (2025) utilized Bloom' s Taxonomy to examine the coverage of higher-order cognitive abilities in existing Large Language Model (LLM) benchmarks. Their findings indicate that current LLMs are biased toward "Remembering"and "Understanding,"with significantly less emphasis on higher-order cognitive requirements, particularly content categorized as "Metacognitive Knowledge" [**?**, **?**].

In this context, the concept of metacognition refers to the knowledge dimension within educational taxonomy—specifically, whether an individual understands how they learn, when they are prone to errors, and which strategies should be employed under specific task conditions [**?**, **?**]. Although this definition is not entirely equivalent to the operationalized concepts of monitoring, evaluating, and regulating one' s own cognitive states found in psychology, it can be viewed as a narrow subset of psychological metacognition focused on knowledge representation. To some extent, these findings demonstrate the conceptual limitations of the current evaluation ecosystem.

2.2 Research on the Classical Metacognitive Paradigm Based on Confidence

In recent years, research on metacognition has increasingly focused on the classical paradigm based on confidence judgments. This approach typically requires participants to perform a primary perceptual or cognitive task (Type 1 task) and subsequently provide a confidence rating (Type 2 judgment) regarding the accuracy of their initial response. This framework allows researchers to quantify the relationship between objective performance and subjective awareness.

Recent studies have refined this paradigm by distinguishing between different components of metacognitive ability, specifically metacognitive bias and metacognitive sensitivity. Metacognitive bias refers to a participant' s overall tendency to report high or low confidence regardless of actual performance, whereas metacognitive sensitivity reflects the degree to which an individual' s confidence ratings can successfully discriminate between their own correct and

incorrect trials. To accurately measure these constructs, researchers have increasingly adopted Signal Detection Theory (SDT) frameworks, particularly the $meta-d'$ metric. This metric quantifies metacognitive sensitivity in a way that is theoretically independent of Type 1 performance (d'), providing a more robust measure of an individual' s "metacognitive efficiency" (the ratio $meta-d'/d'$).

Furthermore, contemporary research has explored the neural correlates and computational mechanisms underlying these confidence-based judgments. Evidence suggests that the prefrontal cortex, particularly the anterior prefrontal cortex (aPFC), plays a critical role in integrating internal signals to form confidence estimates. Computational models, such as the drift-diffusion model (DDM), have been extended to account for the temporal dynamics of confidence formation, suggesting that confidence is often based on the continued accumulation of evidence even after a primary decision has been made. These advancements in the classical paradigm continue to provide deep insights into how the human brain monitors its own internal states and decision-making processes.

Evidence most closely aligned with the classical psychological paradigm of metacognition primarily stems from behavioral regulation based on confidence levels. [**?**] investigated whether models can represent their own numerical confidence levels and utilize these internal signals to regulate second-order decisions, such as the choice to "answer or opt-out." The testing tasks were adapted from paradigms such as post-decision wagering, the opt-out paradigm, and uncertainty monitoring.

The researchers evaluated several models, including GPT-4o, Gemma 3 27B, DeepSeek 671B, and Qwen 80B. Through internal manipulations, they discovered that the confidence levels expressed by Large Language Models (LLMs) are not only correlated with opt-out behavior but also exert a certain causal effect. This suggests that the second-order behavioral regulation observed in these models is not merely a form of post-hoc linguistic packaging; rather, it likely relies on a relatively stable internal mechanism.

internal signals (Kumaran et al., 2026). This positions the evidentiary strength of this study beyond the majority of existing literature, which often remains limited to measuring the correlation between model confidence and accuracy. Instead, this work further demonstrates the causal influence of confidence on accuracy through experimental manipulation. However, the researchers also acknowledge that the evidence obtained here is functional in nature, pertaining to the behavioral and computational levels, rather than a direct demonstration of metacognitive mechanisms in the human sense (Kumaran et al., 2026).

Ackerman (2026) represents an alternative approach to adapting the uncertainty monitoring paradigm for Large Language Models (LLMs). This study reformulates the classic opt-out task—traditionally structured as a "refusal to answer"—into a delegation decision task: "answer the question yourself" versus "delegate the answer to a teammate." While this modification aligns more closely with real-world decision-making scenarios, it simultaneously introduces additional

variables such as modeling the capabilities of others, situational understanding, and cost-benefit trade-offs. Consequently, the conceptual boundaries of what is actually being measured become more blurred.

The researchers evaluated several mainstream and near-frontier models released since 2024, including various models and modalities from Anthropic, OpenAI, Google DeepMind, xAI, DeepSeek, and Alibaba. The findings indicate that some frontier models adjust their willingness to cede answering rights based on internal signals related to their own risk of error. However, these effects are generally weak and exhibit poor stability across different models and materials. Furthermore, surface-level difficulty cues in the questions continue to exert a significant influence on the models' decision-making processes (Ackerman, 2026).

When considered alongside previous research, these results suggest that the internal monitoring signals identified in classic metacognitive paradigms do not easily generalize to task scenarios that more closely resemble everyday decision-making. In such contexts, models appear more susceptible to external cues and the specific framing of the task.

Compared to the previous two studies, Wang et al. (2025) also attempted to distinguish between first-order task performance and second-order control, but they approached the problem through statistical methodology. To facilitate their statistical analysis, the researchers first transformed various benchmark tasks into forced-choice (two-alternative) tests. After obtaining the models' basic confidence and accuracy distributions, they employed a Signal Detection Theory (SDT) framework to separately model and analyze first-order response performance and second-order monitoring capabilities. This approach allowed the researchers to differentiate between task discrimination ability and self-monitoring ability, enabling an investigation into the structural relationship between confidence and accuracy rather than relying solely on simple correlations.

After comparing three mainstream Large Language Models (LLMs)—LLaMA2-70B, GPT-3.5, and GPT-4—the researchers discovered that different confidence elicitation methods systematically altered the experimental results. These findings suggest that whether an LLM "knows" it might be making an error depends heavily on how the researcher externalizes that internal confidence.

(G. Wang et al., 2025). In terms of lateral comparisons, models with stronger overall performance typically exhibit superior metacognitive monitoring capabilities; furthermore, this result remains stable across different methods of eliciting confidence (G. Wang et al., 2025). While such methods—which distinguish between the two stages through statistical analysis—provide slightly less robust evidence than direct causal manipulation, they offer an alternative perspective for observing the implicit metacognitive processes within Large Language Models (LLMs).

Metacognitive Paradigms with Limited Transferability: JOL

Huff and Ulakci (2025) attempted to apply the Judgment of Learning (JOL) paradigm to Large Language Models (LLMs), yet they achieved only partial transferability. Originally, JOL served as a classic prospective monitoring paradigm within metamemory research, designed to investigate whether individuals can accurately predict their future memory performance [**?**]. However, in this specific study, the task assigned to the LLM was not to predict its own internal states or future performance.

In the future, it is worth investigating whether machine learning models can independently score the memorability of experimental materials alongside human participants, rather than merely memorizing the materials themselves. Furthermore, research should examine whether these machine-generated scores can predict subsequent human memory performance as effectively as human Judgments of Learning (JOLs).

Researchers tested three models—GPT-3.5-turbo, GPT-4-turbo, and GPT-4o —via the official OpenAI API. The results indicate a coupling between the Judgments of Learning (JOLs) provided by the human group and their actual memory performance; however, the corresponding ratings generated by current mainstream GPT models failed to predict human performance \cite{Huff_{{{Ulakci}}{{2025}}}}. *The researchers further extended these findings to the fields of education and human-computer interaction, suggesting that enhancing the self-monitoring capabilities of models may hold significant practical value \cite{Huff*{{{Ulakci}}_{{2025}}}}.

However, in these types of tasks, LLMs are essentially performing external predictions rather than monitoring their own future memory performance. Consequently, the conceptual alignment for transferring Judgments of Learning (JOL) to LLMs remains immature, making it difficult to implement true prospective metamemory monitoring within the models themselves.

4. Internal Mechanisms of Metacognition Based on Neurofeedback Research

Compared to the behavioral studies mentioned above, the work of Ji-An et al. (2025) moves closer to a mechanistic exploration of the field. Inspired by the neurofeedback paradigm, this study explores metacognitive mechanisms by mapping internal activations to feedback signals (Ji-An et al., 2025). The researchers primarily evaluated mainstream open-source models following instruction tuning, including various parameter scales of the LLaMA 3 and Qwen 2.5 series. Their results indicate that these models demonstrate a limited but stable second-order mastery over their internal states (Ji-An et al., 2025).

Specifically, the researchers observed that models are capable of learning correspondences between internal activations and external labels, and can even regulate target activation directions to a certain extent. However, this capabil-

ity is primarily concentrated on signal directions characterized by high semantic interpretability and stable statistical structures. Furthermore, this regulatory ability becomes more pronounced as model depth and scale increase (Ji-An et al., 2025). These findings suggest that such regulation remains local and incomplete, rather than representing a universal monitoring mechanism applicable to complex cognitive activities. Although this study employs engineering-based debugging rather than classical metacognitive paradigms as its primary methodology, it provides valuable insights into the internal dynamics of model self-awareness.

However, in terms of conceptual research, they explicitly cite a substantial body of prior literature grounded in psychology. Through a uniquely engineering-oriented perspective, they have successfully demonstrated the inherent limitations of the internal metacognitive mechanisms within current Large Language Models (LLMs).

4.3 工作记忆

Haznitrama et al. (2026) and De Langis et al. (2026) both attempted to migrate comprehensive cognitive measurement protocols to Large Language Models (LLMs). The former incorporated Spatial Working Memory tasks while migrating a neuropsychological assessment framework, while the latter systematically measured forward digit span, backward digit span, operation span, and reading span.

These tasks, including n-back, were utilized to cover various components of working memory, such as short-term retention, simultaneous processing, and continuous updating (De Langis et al., 2026; Haznitrama et al., 2026).

Haznitrama et al. (2026) constructed a cognitive testing framework by adapting Raven' s Advanced Progressive Matrices, Spatial Working Memory tasks, and the Wisconsin Card Sorting Test. They adapted the Spatial Working Memory task into a multi-round search game where the model opens boxes round by round to find target words through a process of elimination. The researchers provided three testing formats: text-only, image-only, and a combination of text and images. The study primarily examined mainstream multimodal reasoning models, including GPT-5, Gemini 3 Pro, Gemini 2.5 Pro, o4-Mini, Claude Sonnet 4, Grok 4/4.1 Fast, as well as GLM 4.5V/4.6V and Qwen3-VL-235B. The results indicated that under simple text-only conditions (Text-easy), most models achieved near-perfect scores. However, performance declined sharply when faced with difficult (Hard) conditions or when image-based and multimodal formats were introduced (Haznitrama et al., 2026). The researchers attributed this to the models' difficulty in maintaining long-range state tracking amidst sparse feedback, as well as logical biases during task execution, such as erroneously selecting non-existent boxes or falling into invalid search loops (Haznitrama et al., 2026). These multimodal working memory tasks reveal significant bottlenecks in current models regarding memory state updating and the integrated application of visual information.

In contrast, De Langis et al. (2026) conducted testing that followed psychological architectures more strictly. The researchers tested six open-source LLMs across three families—Gemma 2, Llama 3.1, and Qwen 2—each including a large and a small version. They also tested two reasoning model variants, including DeepSeek-R1 distilled Llama 3.1-8B and the Qwen 3 series. The results showed that LLMs exceeded human normative levels on most working memory tasks, yet this advantage did not translate into stronger executive control performance (De Langis et al., 2026). The researchers reported that model performance on n-back tasks was close to human levels; on other working memory tasks, particularly simple span tasks, models often reached or exceeded human baselines. Furthermore, parameter scale was significantly positively correlated with average task performance (De Langis et al., 2026). Notably, the models performed almost perfectly on forward digit span tasks even with extremely long sequences, whereas in backward digit span tasks, errors increased significantly as soon as

reverse-order operations were required. Based on this, the researchers concluded that LLMs are strong in information retention and replication but weak in simultaneous processing, updating, and manipulation (De Langis et al., 2026). Synthesizing these findings with the previous study, it is evident that LLM working memory in both text and image modalities only reaches the level of information maintenance; there remain clear deficiencies in the information manipulation and dynamic updating that truly characterize the "working" component of memory.

4.4 记忆监控与元认知领域总体评价

Current evidence primarily supports the observation that Large Language Models (LLMs) exhibit limited metacognitive behaviors, yet it remains difficult to prove that they possess a complete mnemonic mental system. Existing research indicates that in certain metacognitive tasks, models are capable of making specific adjustments—such as deciding whether to continue answering, delegating to others, or modifying their responses—based on uncertainty signals [**?**, **?**, **?**].

Furthermore, LLMs frequently demonstrate strong performance in tasks involving short-term retention and sequential maintenance within working memory [**?**, **?**, **?**]. At the same time, a core issue in this field is the lack of conceptual validity: when classical psychological paradigms are adapted into LLM tasks, confounding factors such as strategy selection, item difficulty, instruction following, and task framing are often introduced, causing a shift in the actual object of study. Future research should move beyond the mere expansion of benchmarks and instead focus on improving conceptual alignment. Priority should be given to advancing research on source monitoring, more rigorously distinguishing between different types of metacognitive evidence, and subdividing the various components of working memory.

5.1 研究范围界定

In psychology, reasoning generally refers to the cognitive process by which individuals generate conclusions, explanations, predictions, or new hypotheses based on existing premises, evidence, rules, relations, or observations. Its core characteristic lies in the transformation, completion, or evaluation of information that is not directly provided, starting from known inputs. Different theoretical traditions emphasize various aspects of the reasoning process: mental model theory highlights the internal representation of premise situations and the search for counterexamples [**?**, **?**]; dual-process approaches focus on the relationship between intuitive responses and analytical processing, particularly regarding belief bias and content effects common in human deductive reasoning \cite{St.

B.

T. Evans et al., 1983}; and research on analogy emphasizes the abstraction of relational structures and cross-domain mapping [**?**, **?**, **?**].

The reasoning discussed in this section is not equivalent to semantic understanding, mathematical calculation, or general problem-solving. To avoid conflating all complex task performance with general reasoning ability, this chapter primarily focuses on the performance of Large Language Models (LLMs) within psychological reasoning paradigms or their adapted tasks. The focus includes deductive reasoning, inductive reasoning, analogical reasoning, and causal reasoning.

5.2 综述与方法学研究

Before proceeding to the empirical study, we first introduce a review and methodological evaluation that guided our research. Mondorf and Plank (2024) argue that previous evaluations of the reasoning capabilities of Large Language Models (LLMs) have relied too heavily on simple accuracy metrics, while rarely analyzing how the models arrive at their answers. They contend that a clear distinction must be made between a model' s final output and the behaviors and internal mechanisms it exhibits when confronted with reasoning tasks (Mondorf & Plank, 2024).

This provides important methodological insights for our evaluation: reasoning assessment must simultaneously focus on both the outcome level and the behavioral level. The outcome level concerns the correctness of the answer, while the behavioral level examines whether the model consistently utilizes task-relevant information, whether its errors are concentrated in specific areas, whether prompting methods alter its judgment, and whether its explanations are consistent with its final answers.

The authors also elaborate on the Chain of Thought (CoT) approach, which prompts models to generate intermediate reasoning steps. They argue that

While requiring a model to report its thinking process can sometimes improve

task performance, these generated texts are not equivalent to the actual internal reasoning process (Mondorf & Plank, 2024). Consequently, we treat CoT as a specific prompting condition to observe its impact on reasoning performance, rather than regarding the CoT output as a direct representation of the LLM' s internal reasoning mechanism.

5.3 演绎推理

In psychology, deductive reasoning refers to the process by which an individual starts from given rules to derive conclusions that are logically necessary. Typical tasks require participants to judge whether a conclusion follows validly from premises, or to select information from several options that can verify, falsify, or complete a rule. Syllogistic reasoning, the Wason selection task, conditional reasoning, propositional logic reasoning, transitive reasoning, and quantifier and negation reasoning all belong to the important paradigms of this tradition [**?**, **?**, **?**]. This is also the subfield of reasoning with the most substantial evidence in current research on the cognitive abilities of Large Language Models (LLMs).

(1) Direct Transfer Research of Classical Deductive Paradigms

Eisape et al. (2024) directly transferred the classical syllogism paradigm to language model evaluation, using 64 types of standard syllogistic structures to compare four models of varying sizes from the PaLM 2 family with human behavioral data.

The results showed that larger model scales correlated with better performance, with some models exceeding human average accuracy; however, even the best-performing model achieved an accuracy of only 75% [**?**, **?**]. Simultaneously, the models exhibited human-like reasoning biases, being influenced by variable order and specific syllogistic structures in patterns similar to those seen in humans. Notably, when faced with problems designed such that no conclusion could be drawn, the models were often reluctant to admit that the premises were insufficient to support a conclusion [**?**, **?**]. As an early systematic measurement of classical syllogistic paradigms in LLMs, the conclusions of this study carry significant weight.

The reasoning biases and blind spots presented by the models, which mirror those of humans, suggest at the very least that cognitive psychology theories can be utilized to explain certain behaviors of LLMs.

Ozeki et al. (2024) similarly investigated the performance of GPT-3.5, GPT-4, Llama-2 (13B / 70B), and Swallow (13B / 70B) across different conditional syllogism tasks. Using the NeuBAROCO dataset, researchers required the models to judge the relationship between two premises and a conclusion. In some items, the conclusion could be derived from the premises; in others, the conclusion contradicted the premises; and in some, the premises were insufficient to support any definitive judgment.

To examine the influence of semantic content, the researchers also designed different types of materials, including abstract symbolic materials, natural language materials consistent with common sense, and natural language materials inconsistent with common sense.

The results indicated that GPT-4 achieved the best overall performance, and increases in model scale were accompanied by improvements in accuracy. However, similar to previous findings, the models were more prone to errors on items where the premises were insufficient for judgment, and accuracy decreased when dealing with materials inconsistent with common sense [**?**, **?**]. It is worth noting that models sometimes demonstrated a good understanding of the natural language premises in the prompts yet still failed at the final logical judgment. Based on this, the researchers concluded that the errors made by models in deductive reasoning primarily stem from the reasoning stage rather than the language comprehension stage [**?**, **?**]. This research design—manipulating linguistic materials to distinguish between semantic comprehension errors and reasoning errors—goes beyond surface-level accuracy measurements to deconstruct the internal patterns of LLMs in syllogistic tasks.

Lampinen et al. (2024) compared the performance of humans and LLMs in syllogisms, natural language inference, and Wason selection tasks. Testing Chinchilla, PaLM 2-M, PaLM 2-L, Flan-PaLM 2, and GPT-3.5-turbo-instruct, they found that both models and humans were more likely to accept conclusions that were believable or consistent with realistic experience. In the Wason task, model performance under realistic rule conditions was generally superior to that under arbitrary rule conditions [**?**, **?**]. Stone et al. (2024) and Abe et al. (2026) further demonstrated that LLMs find it easier to complete the parts of the Wason selection task involving direct rule verification, while performance drops in parts requiring the active search for counterexamples. Furthermore, models typically perform better on deontic rules involving social norms than on abstract descriptive rules. Models are also frequently misled by the surface terms of a problem; for instance, they tend to select terms that appear directly in the rules rather than consistently judging according to the correct logic of seeking counterexamples [**?**, **?**, **?**]. Taken together, LLMs clearly exhibit human-like content effects in behavioral performance during conditional reasoning and syllogistic tasks. This suggests that LLMs do not strictly follow content-independent formal reasoning when performing deductive tasks but, like humans, often rely on real-world knowledge and linguistic cues within the prompt.

Unlike the categorical inclusion relationships common in syllogisms, transitive reasoning typically requires individuals to integrate information from premises to find patterns. For example, if the premises inform the model that "A is higher than B," "B is higher than C," and "C is higher than D," the model must further infer relationships not directly provided, such as "A is higher than C" or "A is higher than D."

Wu and Deng (2025) adapted this classical paradigm into a natural language task, requiring GPT-3.5-Turbo, GPT-4, Llama3-8B, and Qwen to judge whether

new comparative relationships held based on paired information such as student grades, food preferences, or employee salaries. The results showed that all four models performed above chance levels, with GPT-4 and Llama3-8B performing particularly well [**?**, **?**]. Meanwhile, model performance was significantly influenced by the organization of the material: when premises were presented in a continuous sequential order, the performance of GPT-3.5-Turbo, GPT-4, and Llama3-8B was superior to their performance when the order of premises was scrambled. Some models also exhibited behavioral patterns similar to those found in human transitive reasoning research, such as finding it easier to judge comparisons involving the highest or lowest items, as well as pairs with larger differences [**?**, **?**].

Overall, LLMs can achieve high performance in syllogisms, Wason selection tasks, and transitive reasoning, exhibiting many human-like behavioral patterns and biases. However, these similarities remain primarily at the behavioral level. The numerous difficulties models encounter in classical tasks also indicate that the deductive reasoning capability of LLMs is not a stable form of formal reasoning.

(2) Research on the Adaptation and Extension of Classical Deductive Paradigms

Research in this area primarily focuses on modifying and extending traditional syllogistic assessments. Consequently, the questions of interest have shifted: whether a model can correctly answer a syllogism is only the first level; more important is the comparison of the model across different logical structures, its ability to understand and translate natural language into logical forms, and whether external visual representations used by humans are effective for the models. Delgado-Solorzano et al. (2024) and Zong and Lin (2024) both argue that model performance in syllogistic tasks depends largely on whether the test covers a complete set of logical structures. A "complete structure" primarily refers to the combination of different quantifier sentence types and different term order positions—for example, propositional forms such as "All," "Some," "None," and "Some···not," as well as the arrangement of terms A, B, and C in the premises and conclusions. Delgado-Solorzano et al. (2024) evaluated models using a comprehensive set of 256 syllogistic structures, finding that Claude 3 Opus and GPT-4 could outperform the average human on some item types but still made errors on complex, specific logical structures [**?**, **?**].

Zong and Lin (2024) further pointed out that templated syllogism datasets can cover these structures more completely, and models perform better on such materials. In contrast, syllogism datasets described in natural language offer richer expression but typically cover a narrower range of logical structures and introduce issues of semantic comprehension [**?**, **?**]. GPT-4 and GPT-4o showed high accuracy on templated materials, but their performance dropped significantly on human-generated natural language materials. This indicates that the models' difficulties stem not only from logical reasoning itself but also from

the process of abstracting natural language premises into standard syllogistic structures [**?**, **?**].

Another category of extension research attempts to use diagrams to assist deductive reasoning. Ando et al. (2024) investigated whether Euler diagrams could help models better complete syllogistic judgments. Euler diagrams essentially use circles to represent the scope of concepts, allowing the model to represent relationships between different concepts through partially overlapping circles. The results showed that inputting Euler diagrams helped with some valid conclusions, but for items where no valid conclusion could be drawn, the models still committed the over-inference errors mentioned previously [**?**, **?**]. This suggests that while external graphics can help LLMs reduce the burden of representation to some extent, they cannot solve the model' s tendency toward over-inference.

Taken together, adaptation and extension research has shifted the explanatory focus of syllogistic assessment from simple accuracy toward the sources of error. This further demonstrates that the evaluation of deductive reasoning in LLMs must be interpreted in conjunction with task formats and scoring methods.

(3) Practical Tasks Based on Deductive Paradigms

Deductive structures have also been applied to task design in practical scenarios. Song et al. (2025) introduced the legal syllogism structure into legal text summarization, organizing summary generation based on the relationships between legal rules, case facts, and conclusions. After adopting the syllogistic reasoning framework, the model' s performance improved across legal domain benchmarks such as ROUGE-L, BLEU, and BERTScore. This study demonstrates how a deductive reasoning framework can help LLMs output text that better conforms to specific domain norms [**?**, **?**].

5.4 归纳推理

Inductive reasoning refers to an individual' s ability to generalize general laws from exemplary cases through empirical observation. Category induction research examines how individuals infer whether other category members or superordinate categories possess a specific attribute based on its presence in a given category member. Rule discovery tasks are often represented by Wason' s 2-4-6 task, which requires individuals to propose and test rules from a small number of samples. Probability induction research focuses on how individuals make predictions based on limited information and prior probabilities. Finally, sequence or pattern induction tasks examine whether individuals can discover regularities and perform extrapolations from sequences of letters, numbers, or symbols [**?**, **?**, **?**, **?**, **?**, **?**]. Human inductive judgments are frequently influenced by representativeness, similarity, sample size, and prior knowledge [**?**, **?**].

(1) Research on the Direct Transfer of Classical Inductive Reasoning Paradigms Compared to deductive and analogical reasoning, we find that current literature lacks empirical studies on Large Language Models

(LLMs) that independently center on classical psychological inductive reasoning paradigms. Inductive reasoning primarily appears as a component within comprehensive benchmarks.

Consequently, we will introduce these comprehensive reasoning benchmarks in the next section and highlight the performance of inductive reasoning within them. This absence of independent study may be related to the measurement characteristics of inductive reasoning itself; classical inductive tasks typically require examining how individuals generalize rules from limited examples. The answers to such tasks are often not as definitive as those in deductive reasoning, making them more difficult to adapt into automated scoring tests.

(2) Adapted and Extended Inductive Reasoning and Cross-Type Reasoning Benchmarks Due to the current lack of LLM research directly transferring classical inductive reasoning paradigms, comprehensive reasoning benchmarks have become the primary material for observing inductive performance. These benchmarks typically incorporate a mix of reasoning tasks—including deductive, inductive, analogical, and causal reasoning—to compare the distribution of model performance across different reasoning types.

Xu et al. [**?**, **?**] proposed a comprehensive logical reasoning evaluation framework covering deductive, inductive, causal, and mixed reasoning, using fifteen datasets and various error types to analyze model performance. They tested seven models, including LLaMA-3.1-Chat, Mistral-Instruct-v0.3, Claude-3, and GPT-4. The results showed that while LLMs can provide correct answers for some inductive tasks, their performance is highly unstable. Specifically, models may occasionally select the correct answer without necessarily being able to clearly explain the rules they followed. The researchers concluded that a small number of examples often only helps the model familiarize itself with the question format and is insufficient for them to stably abstract general rules [**?**, **?**].

They further discovered significant performance differences across different task formats; tasks requiring the model to generate answers directly are generally more difficult than those requiring classification among multiple options. More importantly, across all tasks, when researchers required the simultaneous presence of a correct answer, a correct explanation, and a complete explanation, the performance of all models dropped significantly. Errors primarily manifested as inaccurate identification of key information in explanations, planning errors in the reasoning logic itself, and disordered execution steps [**?**, **?**]. This suggests that even when models provide correct answers in inductive and other reasoning tasks, they have not necessarily formed a rigorous and complete reasoning process [**?**, **?**].

The LogiEval and LogiEval-Hard benchmarks constructed by Liu et al. [**?**, **?**] provide further comprehensive evidence regarding reasoning. This study evaluated the performance of Claude 3.7 Sonnet Thinking, DeepSeek-R1, Gemini 2.0 Flash Thinking, Grok3 Think, OpenAI o4-mini, Qwen3-235B-A22B, and QwQ

32B on deductive, inductive, analogical, and causal reasoning questions from human examinations. The average accuracy of the models on standard difficulty levels concentrated between 78.74% and 81.41% [?, ?]. Because the study did not strictly adopt psychological paradigms but instead drew materials from logical reasoning questions in the Chinese Civil Service Exam and intelligence tests, its orientation is more practical and possesses higher ecological validity.

Regarding induction-related tasks, the models' strengths are mainly reflected in abstract pattern recognition tasks, while tasks requiring multi-step reasoning, such as kinship relations, remain difficult to score [?, ?]. Liu et al. [?, ?] also found that the distribution of task difficulty for models does not align with that of humans: models can solve some problems that humans find difficult, yet perform poorly on problems of medium difficulty for humans [?, ?].

This indicates that the high overall scores achieved by models cannot yet be interpreted as the models possessing the same reasoning structures as humans.

Taken together, these two studies provide only indirect evidence for the inductive reasoning capabilities of LLMs. This is primarily because the "inductive reasoning" within these comprehensive benchmarks makes it difficult to distinguish whether a model is truly generalizing rules or merely utilizing local cues within the material to complete the task. Therefore, we can only state that LLMs demonstrate a certain ability to complete inductive reasoning tasks within comprehensive logical benchmarks, but this is insufficient to prove they possess stable inductive reasoning capabilities in the psychological sense.

5.5 类比推理

In psychology, analogical reasoning focuses on an individual' s ability to identify shared relational structures between two distinct contexts and transfer these analogical relationships to other domains with similar structures. For example, two stories may feature different characters on the surface, yet share the same underlying plot dynamics; similarly, two geometric matrices may differ in visual appearance while following the same fundamental rules of transformation.

Analogical reasoning research frequently employs a variety of task formats, including letter-string analogies, verbal analogies, story analogies, Raven' s Progressive Matrices, and structure-mapping tasks. Letter-string analogies are common in the tradition of computational cognitive models such as Copycat; these tasks require participants or models to complete an analogy based on the transformational relationships between sequences of letters. Verbal analogies are often used to investigate the information-processing components of analogical reasoning, while story analogies examine whether individuals can transfer the relational structure of one narrative to another problem context with different surface features. Finally, Raven' s Progressive Matrices and visual matrix tasks are standard measures for assessing abstract relational reasoning and fluid intelligence [?, ?, ?, ?, ?, ?].

Direct Transfer Studies of Classical Analogy Paradigms

Musker et al. (2025) conducted a comparative study between Large Language Models (LLMs) and humans across a suite of analogy tasks. Their results demonstrated that under standard conditions—characterized by clear materials and stable formatting—models such as GPT-4, Claude 3, and Llama-405B can approach human-level performance, successfully executing a degree of relational mapping [**?**]. However, when the researchers modified the presentation of the materials—for instance, by shuffling the order of items or introducing word structures irrelevant to the correct answer—they observed a significant decline in model performance. Notably, the models proved to be more susceptible to the influence of irrelevant linguistic distractors than human participants [**?**].

To further investigate these capabilities, the researchers altered the task formats to minimize the possibility of models completing answers based solely on fixed structural patterns. They found that the high scores previously achieved by several models likely depended on specific material formatting; among the tested models, only Claude 3 remained relatively close to human performance under these controlled conditions [**?**]. These findings suggest that while mainstream LLMs have developed a foundational capacity for analogical relational mapping, this ability in many models remains heavily dependent on how materials are organized and is easily disrupted by irrelevant semantic information.

Haznitrama et al. (2026) investigated the abstract relational reasoning capabilities of models using Raven' s Progressive Matrices (RPM) tasks. The RPM task typically presents a matrix of geometric figures with one missing cell; the model must identify the most appropriate completion from a set of candidates by discerning how the shapes, quantities, positions, or orientations of the images change across rows and columns. The researchers implemented both text-based versions (where each figure is described verbally) and image-based versions of the Raven tasks. Their results indicated that models generally performed better under text-based conditions than image-based conditions. Furthermore, in the image-based tasks, all models performed significantly below the human benchmark; even the top-performing model, Qwen3-VL-235B, failed to reach the average human level of performance [**?**].

Researchers have also discovered that requiring a model to utilize Chain-of-Thought (CoT) prompting—where the model articulates its reasoning process —is not always beneficial. Specifically, in certain multiple-choice text tasks, models actually perform better when they are not required to provide step-by-step reasoning [**?**, **?**]. This research suggests that because Raven-style tasks are primarily presented in visual formats, they evaluate more than just the Large Language Model' s (LLM) analogical reasoning capabilities; they also serve as a significant test of the model' s ability to recognize and interpret images.

This challenge is further reflected in subsequent studies that attempt to adapt Raven-style tasks for LLMs. Currently, there is a lack of research designs capable of isolating this confounding variable, making it difficult to directly and

cleanly migrate Raven-style tasks to LLMs for the sole purpose of measuring pure reasoning ability.

2.2 Research on Adapting and Extending Classical Paradigms of Analogical Reasoning

In the field of adaptation studies concerning analogical reasoning, the most significant contributions involve new paradigms that test model stability by introducing minor perturbations to the original problems.

A representative example of this approach is the work of Webb et al. [**?**], who developed a zero-shot reasoning task based on the Digit Matrix (a simplified version of Raven' s Progressive Matrices). In this task, the model must identify the underlying transformation rule within a matrix of numbers and apply it to complete the pattern. To rigorously evaluate whether Large Language Models (LLMs) truly grasp the logic of the task rather than relying on memorized patterns, the researchers introduced a "counter-intuitive" perturbation method. By altering the standard numerical order or changing the labels of the categories, they created scenarios that contradicted common human intuition. Their results demonstrated that while models like GPT-3.5 and GPT-4 showed strong performance on standard tasks, their accuracy significantly declined under these perturbed conditions, suggesting a lack of robust, generalized reasoning capabilities.

Furthermore, recent studies have extended these paradigms to include cross-modal and multi-step analogical reasoning. These extensions require models to not only recognize simple $A : B :: C : D$ relationships but also to maintain consistency across different representational formats or through complex, sequential logical chains. By systematically varying the complexity and the presentation of these analogies, researchers can better delineate the boundaries between superficial pattern matching and genuine structural mapping in artificial intelligence.

Lewis and Mitchell (2024) conducted a robustness test regarding the conclusions drawn by Webb et al. (2023), who suggested that GPT-3 possesses emergent analogical reasoning capabilities that do not require additional prompting. Directly adopting the three task categories used by Webb et al. (2023)—letter string analogies, Raven' s Progressive Matrices (digit-based), and story analogies—the researchers constructed variant problems that preserved the underlying abstract rules while altering the surface forms. By comparing the performance shifts of both humans and the GPT series models across the original and variant problems, they aimed to evaluate the stability of these reasoning abilities (Lewis & Mitchell, 2024).

The results indicate that humans exhibit greater resilience to interference than models in simple letter-string tasks. In the Raven' s Progressive Matrices (digit-based), GPT-3.5 and GPT-4 maintained their performance levels when symbols were replaced; however, their performance declined significantly when the position requiring completion began to vary randomly. Furthermore, in story-

based analogy tasks, GPT-4 demonstrated a pronounced order effect [**?**, **?**]. These findings provide strong evidence that Large Language Models (LLMs) are easily influenced by variations in problem phrasing and format during analogy tasks, remaining far from the stable performance characteristic of humans. Consequently, robustness testing is essential for validating the true analogical capabilities of LLMs.

Beyond robustness testing, the adaptation of analogical reasoning tasks is more diverse compared to other subfields of machine learning. Combs et al. (2025) adapted the story analogy task from psychology. Traditional story analogy research focuses on whether individuals can identify structural similarities between two stories.

Analogical reasoning involves identifying identical deep relational structures between different events or stories (Gick & Holyoak, 1980). Combs et al. (2025) extended this concept into a long-text zero-shot comparison task, requiring models to determine which of several natural language stories shares the same plot structure as a target story. Their results indicated that while GPT-4 and GPT-4o performed well on certain datasets, multiple models performed near or below chance levels on more challenging datasets. Further analysis revealed that these models are easily distracted by similarities in characters, settings, and specific entities; consequently, they exhibit instability when identifying stories that share identical plot relations but differ in surface-level content (Combs et al., 2025).

Webb et al. (2025) adapted the letter-string analogy task to further investigate model capabilities. Traditional letter-string analogies typically require subjects to complete a sequence based on relational changes between letters, such as identifying rules like "move one position forward" or "repeat a specific position" [**?**]. In their study, the researchers replaced the conventional alphabet with a randomized, hypothetical alphabet. This modification ensures that models cannot rely on the familiar A, B, C sequence and must instead complete the analogy based on the novel ordering provided within the prompt [**?**].

The results indicated that GPT-4 performed significantly worse than humans under these conditions, with Chain-of-Thought (CoT) prompting providing only limited improvement. While the model' s performance approached human levels when it was permitted to call and execute code, this improvement was primarily driven by external computational tools. Consequently, these gains cannot be directly interpreted as an enhancement of the model' s inherent analogical reasoning capabilities within a natural language framework [**?**].

Camposampiero et al. (2025) adapted the Raven' s Progressive Matrices task for their study. While the traditional Raven task requires subjects to complete a missing entry based on the underlying patterns of geometric figures in a matrix [**?**], the researchers transformed this into a symbolic matrix task. In this version, the original graphical matrices are converted into matrices composed of symbolic attributes such as shape, color, quantity, and position. Models are then required to complete the missing entry by identifying the governing rules

of these attributes [?].

To further challenge the models, the researchers introduced enhanced interference mechanisms. For instance, they included variations in attributes that are irrelevant to the correct answer, thereby hindering the model's ability to identify the true underlying rules. Additionally, they made certain symbolic attributes ambiguous or continuous, forcing the models to discern stable patterns amidst fuzzy information [?]. The results demonstrated that while OpenAI o3-mini and DeepSeek R1 performed strongly under standard symbolic matrix conditions, their accuracy declined sharply when irrelevant attributes or uncertain cues were introduced. Notably, artificially increasing the models' reasoning time did not consistently recover their performance.

Zhang (2024) also draws inspiration from Raven' s Progressive Matrices and visual combinatorial reasoning tasks, primarily examining how subjects identify the ways in which graphical elements are combined, moved, or transformed. They adapted these visual analogy tasks for the evaluation of Multimodal Large Language Models (MLLMs). Compared to traditional Raven-style tasks, this approach places greater emphasis on the model' s ability to accurately identify fundamental information within an image, such as shapes, quantities, and positions [?, ?].

After testing Qwen2-VL, LLaVA 1.6, and LLaMA 3.2, the researchers found that the models performed poorly when no reasoning prompts were provided; however, performance improved after incorporating a single reasoning prompt. A detailed case analysis further revealed that many of the models' errors did not occur during the final stage of relational reasoning. Instead, they occurred during the initial visual recognition stage: the models failed to correctly describe the basic graphical elements within the matrix, which subsequently led to errors in rule induction and answer selection [?, ?]. Based on this, we can infer that in current multimodal models, the primary bottleneck often lies in the foundational perception of visual details rather than the high-level logical processing itself.

Until image recognition capabilities reach a certain stage of development, the validity of reasoning tasks involving images cannot truly align with the underlying reasoning ability itself, unless an experimental design can be found that losslessly isolates the impact of visual processing.

Taken as a whole, the value of adapted analogy tasks lies in testing whether a model' s analogical performance can transcend fixed formats and truly move beyond surface-level similarities or reliance on external tools. Current results indicate that the performance and robustness of existing models in these types of tasks remain insufficient [?, ?, ?, ?].

5.6 因果推理

Causal reasoning refers to an individual's ability to identify, explain, predict, and make intervention-based judgments regarding the causal relationships between

events. Compared to general associative judgment, causal reasoning requires an individual to determine whether an outcome would change if a specific operation or manipulation were performed on the system.

Pearl (2009) distinguished causal reasoning into three hierarchical levels: observation, intervention, and counterfactuals [**?**]. This framework has provided an essential reference for the evaluation of causal reasoning in Large Language Models (LLMs) in recent years.

(1) Research on the Direct Transfer of Classical Causal Frameworks Wang and Shen (2024) designed three types of textual situational tasks based on Pearl' s observation, intervention, and counterfactual framework. The first category requires the model to determine whether one event causes another based on a given causal structure. The second category requires the model to judge whether the original causal relationship still holds after an external human intervention. The third category utilizes Simpson' s Paradox scenarios to examine whether the model can identify which judgment more accurately aligns with a causal explanation when aggregate trends conflict with individual trends [**?**].

The results indicated that among the eight models tested, GPT-4 performed best in most tasks; however, the performance of all models was unstable, particularly in the Simpson' s Paradox task. Intervention conditions significantly increased task difficulty; while models might successfully complete reasoning under an original causal chain, they often failed to adjust their judgments accordingly when external interventions altered the causal path [**?**]. Similar to previous findings, the researchers also observed that models could sometimes select the correct answer while providing incorrect or incomplete causal explanations [**?**].

Consequently, we can only conclude that LLMs demonstrate a certain degree of causal judgment capability in simple scenarios. However, they still face significant difficulties under interventionist and complex mixed conditions, and their reasoning outcomes lack a stable correspondence with their generated explanations.

(2) Research on Adapted and Extended Causal Reasoning

Abe et al. (2024) investigated the abductive reasoning of LLMs. Abductive reasoning examines the ability to reason backward from an outcome to a possible explanation; unlike deductive reasoning, it requires inferring plausible possibilities rather than necessary conclusions. The researchers adapted the traditional syllogism into a "rule-observation-hypothesis" tripartite task, asking the model to judge whether a specific hypothesis serves as a reasonable explanation based on a rule and an observation. Some hypotheses were reasonable, some were unreasonable, and in other cases, the provided information was insufficient to determine the hypothesis' s suitability [**?**].

The results showed that LLMs find these abductive tasks significantly more difficult than original deductive reasoning tasks. In zero-shot conditions, even

the best-performing model, GPT-4, performed only slightly above the random chance level for a three-choice task. Furthermore, the models struggled most with items where the provided information was insufficient to make a judgment, often forcing a definitive affirmative or negative conclusion [**?**]. The study also found that models were more prone to errors when the task content conflicted with everyday common sense. Additionally, when negative expressions such as "no" or "not" appeared in the prompt, the models were more likely to be influenced by the negative form and select a hypothesis containing a negation.

It is evident that in both deductive tasks and adapted abductive tasks, LLMs tend to avoid uncertain answers and lean toward providing seemingly definitive judgments. This suggests that accurately identifying the boundaries of a conclusion remains a common weakness in model reasoning.

5.7 推理领域总体评价

Overall, compared to many comprehensive ability assessments, the advantage of psychological reasoning paradigms lies in their capacity to decompose model performance into richer, more granular levels. These include whether a model can understand premises, abstract relationships, resist semantic interference, handle counterexamples, and cease inference when evidence is insufficient. Existing research collectively indicates that Large Language Models (LLMs) have already demonstrated reasoning-like behaviors in certain deductive and inductive tasks, as well as in analogy and causal judgment. However, this capability is highly dependent on task formatting, material phrasing, prompting methods, and the available answer space.

The most notable common finding in this field is that LLMs often struggle to grasp the boundaries of reasoning. In deductive tasks, this manifests as a reluctance to acknowledge when a conclusion cannot be derived from the given premises (Eisape et al., 2024). In inductive tasks, it appears as the ability to utilize local cues to provide answers without necessarily stabilizing the underlying generalized rules (F. Xu et al., 2025). In analogy tasks, models frequently mistake surface-level similarity for deep structural relationships (Musker et al., 2025). Finally, in causal and abductive tasks, models tend to provide explanations that are superficially plausible but lack sufficient evidentiary support (L. Wang & Shen, 2024).

Consequently, we argue that the current reasoning capabilities of LLMs should be understood as a conditional, explicit performance. While this currently supports assessments of capability at the behavioral level, there remains a significant lack of research regarding the stability and underlying mechanisms of these processes.

6.1 研究范围界定

In psychology, problem-solving is generally understood as a cognitive process in which an individual, faced with an obstacle between their current state and a target state, achieves that goal by forming a problem representation, searching for possible paths, and selecting and adjusting strategies (Mayer, 1992; Newell et al., 1972). Early Gestalt psychology emphasized the restructuring of and insight into problems, suggesting that individuals must change their understanding of a problem' s structure to escape the constraints imposed by their original representation (Duncker & Lees, 1945). Subsequently, the information-processing approach viewed problem-solving as a search process within a problem space, focusing on the relationships between the initial state, the goal state, operators, and intermediate states (Newell et al., 1972).

Within a broader framework of cognitive control, problem-solving also relies on the support of executive functions: working memory helps individuals maintain goals and intermediate information, inhibitory control assists in suppressing salient but incorrect responses, and cognitive flexibility supports rule switching and strategy updating (Diamond, 2013; Miyake et al., 2000). Simultaneously, human problem-solving and decision-making processes are influenced by mental sets, functional fixedness, heuristics, and judgment biases. For instance, familiarity with a specific solution may hinder the discovery of new strategies, while framing, anchoring, and representativeness cues can systematically alter an individual' s judgment (Luchins, 1942; Tversky & Kahneman, 1974; Watson, 2011).

Building upon previous research frameworks, we primarily examine four categories of test paradigms within the domain of problem-solving. The first category involves planning and sub-goal decomposition, focusing on whether a model can formulate goal paths, identify intermediate states, and replan when environmental structures or reward information change. The second category concerns rule switching and cognitive flexibility, focusing on whether a model can discover rules based on feedback, maintain current rules, and abandon old rules once they have changed. The third category addresses mental sets, functional fixedness, and insight-based problem-solving, examining whether a model becomes trapped by familiar cues, default solutions, or misleading information, and whether it can re-represent the problem. The fourth category covers judgment and decision-making biases, focusing on whether a model exhibits cognitive biases—such as anchoring, framing effects, confirmation bias, the representativeness heuristic, and the availability heuristic—when dealing with uncertainty, probability, risk, and value assessment.

6.2 规划与子目标分解

In the field of planning and sub-goal decomposition, the paradigms covered in the current literature remain relatively limited. In classical cognitive psychology, planning typically refers to the process by which an individual searches for a se-

quence of actionable steps between a current state and a goal state, progressively adjusting strategies based on intermediate states. Common measurement tasks include the Tower of Hanoi, the Tower of London, water jug problems, mazes, path planning, means-ends analysis tasks, and multi-step action sequence tasks [**?**, **?**]. These tasks focus on whether an individual can depart from a current state, identify intermediate states, arrange the sequence of actions, avoid obstacles, and adjust plans as conditions change.

However, with the exception of path planning, many of these tasks require subjects to operate continuously within a dynamic state space, monitor intermediate states, and update plans based on the outcomes of their actions. For large language models (LLMs) to fully replicate

these types of tasks, the models must be embedded within interactive environments or agent frameworks. This allows them to execute actions, receive environmental feedback, and continuously update their internal states [**?**, **?**]. Furthermore, path-planning tasks are difficult to apply directly to LLMs because these models can only receive text-based or image-based descriptions and provide routes, choices, or explanations through linguistic output. While many studies utilize various technical means to enable models to simulate human-like path planning in physical or virtual spaces, such research tends toward engineering and essentially explores the internal representation of spatial information in LLMs from a technical perspective.

Because this review focuses primarily on the performance of mainstream LLMs in relatively direct psychological measurement tasks, we ultimately identified only one study that met the inclusion criteria. This study, which employs an adapted version of the classical path-planning paradigm, serves as our primary source of evidence for examining the planning capabilities of LLMs.

(1) Research on Adapted and Extended Planning Paradigms The CogEval framework proposed by Momennejad et al. [**?**] is the core study regarding planning and sub-goal decomposition in the current literature. This research adapts the classical cognitive map paradigm [**?**] by simplifying environmental exploration into textually described locations and connectivity relationships. Models are then required to select routes based on these descriptions and re-select paths when rewards, pathways, or goals change [**?**].

The researchers tested models including GPT-4, GPT-3.5-turbo, Google Bard, Anthropic Claude-1, and LLaMA-13B. The results indicated that in simple path tasks, LLMs can utilize the connectivity relationships directly provided in the text to achieve a certain degree of route selection. However, performance becomes significantly unstable when tasks require the model to perform multi-step path inference or rearrange routes based on changes in rewards or pathways [**?**]. The researchers summarized the error patterns of LLMs, finding that the models frequently hallucinate non-existent pathways and generate circuitous or repetitive routes; even when local connection prompts are provided, the models may

still fail to form a coherent global path [**?**]. From this, we can observe that while LLMs can handle simple path relationships, they struggle to stably maintain more complex path structures and perform flexible planning accordingly.

The scarcity of evidence related to planning and path searching suggests that, compared to other cognitive domains, classical paradigms in problem-solving and executive control are more difficult to migrate to LLM research. We believe this is primarily because classical problem-solving paradigms often require subjects to explore, operate, engage in trial-and-error, receive feedback, and adjust behavior within an environment, whereas current mainstream LLMs respond primarily through text or image inputs. Consequently, to investigate problem-solving processes that more closely resemble reality, researchers must typically convert original paradigms into text-based tasks or introduce external tools and build agent frameworks.

The former approach makes measurement more indirect, while the latter tends to shift the research focus toward the overall performance of the engineering system, rather than purely reflecting the model' s own performance on classical psychological constructs.

Currently, the mainstream way the public uses LLMs remains limited to text and image interaction. However, as multimodal interaction, tool use, and agent systems mature, researchers may be able to design measurement methods that more closely approximate real-world exploration and action processes. At that point, the assessment of LLM problem-solving capabilities within the psychological domain may enter an entirely new frontier.

6.3 规则转换与认知灵活性

Rule switching and cognitive flexibility refer to an individual' s ability to promptly adjust their original thinking and adopt new rules better suited to the current context when task requirements, environmental cues, or feedback results change. Cognitive flexibility is generally regarded as a core component of executive function, working alongside working memory and inhibitory control to support complex problem-solving (Miyake et al., 2000; Diamond, 2013). The most classic measurement paradigm for this construct is the Wisconsin Card Sorting Test (WCST). In this task, participants must infer whether cards should be classified by color, shape, or number based on "correct/incorrect" feedback provided by the experimenter. Once the participant has mastered the current rule, the rule is changed without explicit warning. Researchers then observe whether the participant can cease using the old rule and establish a new classification rule based on the updated feedback (Berg, 1948; Grant & Berg, 1948).

1. Direct Transfer Research within the Classical Rule-Switching Paradigm

Recent studies by Goto et al. (2025), Hao et al. (2025), and Li et al. (2025) have all utilized the Wisconsin Card Sorting Test (WCST) to investigate the rule-switching capabilities of Large Language Models (LLMs), with each study emphasizing different aspects of the transfer process. Goto et al. (2025) adapted the original card-based task into a purely textual format, requiring models to perform classification based on written descriptions of attributes such as color, shape, and quantity. Their analysis focused primarily on the performance disparities between various models regarding rule discovery, rule maintenance, and behavioral adjustment following a rule shift.

Hao et al. (2025) extended this line of inquiry by applying the WCST to Vision-Language Models (VLMs), comparing performance across visual versus textual inputs and direct response versus reasoning-based prompting. This approach highlights how visual modalities and specific prompting strategies influence model performance; notably, the researchers also recruited 30 human participants to establish a comparative human baseline. Meanwhile, Li et al. (2025) integrated the WCST with the Iowa Gambling Task and the Cambridge Gambling Task, situating rule-switching within a broader framework that includes risky decision-making and feedback learning. Their study also incorporated a substantial human baseline consisting of 350 participants to benchmark model performance against human behavior.

All three studies found that powerful models—such as ChatGPT o1, Claude-3.5 Sonnet, Gemini-1.5 Pro, GPT-4o, and o1-mini—demonstrate a degree of rule discovery and switching capability within the clearly structured Wisconsin Card Sorting Test (WCST). Some models even approached or exceeded human baselines in terms of overall accuracy or rule completion speed \cite{Goto_{{{et}}*{{al}}}{2025}, G_{{{Hao}}*{{et}}}{{{al}}*{{2025}}}, *H*{{{Li}}*{{et}}*{{{al}}_{{2025}}}}. At the same time, these studies indicate that task accuracy alone is insufficient to measure a model' s true cognitive flexibility. Specifically, Goto et al. (2025) discovered that while some models are capable of inferring rules···

···can identify the current rule changes but fail to consistently implement these rules in subsequent selections. Hao et al. [**?**] found that models perform significantly better on text-based problems, whereas they are more prone to errors when processing visual inputs. Furthermore, Li et al. [**?**] discovered that the fundamental difference between Large Language Models (LLMs) and humans lies in the nature of their mistakes: while human errors are often sporadic—primarily caused by brief lapses in concentration that lead to temporary deviations from the current rule—LLMs are more likely to exhibit perseverative errors. In these instances, the model continues to categorize according to the outdated rules from a previous stage, even after the rules have clearly changed.

Overall, Large Language Models (LLMs) demonstrate the ability to identify

rules and switch between them to a certain extent within fixed, classical Wisconsin Card Sorting Test (WCST) tasks; however, this capability remains unstable. The performance of these models may deteriorate significantly when presented with original visual test materials, and their adaptation speed during rule transitions is notably slower than that of humans. Consequently, it can be concluded that LLMs do not yet possess the same level of robust cognitive flexibility as humans within classical rule-switching paradigms.

2.2 Research on Adapted and Extended Rule-Switching

Within the scope of our included studies, rule-switching tasks are frequently employed as a component for measuring overall cognitive executive control. The following two articles both argue against using scores from a single type of task to represent foundational cognitive abilities. Instead, they emphasize that analytical significance is only achieved through the simultaneous measurement of multiple functions (De Langis et al., 2026; Haznitrama et al., 2026).

The NeuroCognition battery developed by Haznitrama et al. (2026) comprises three distinct tasks: Raven' s Progressive Matrices, Spatial Search, and the Wisconsin Card Sorting Test (WCST). These tasks are specifically designed to evaluate a model' s capabilities in analogical reasoning, spatial working memory, and rule switching, respectively.

The results of the Wisconsin Card Sorting Test (WCST) indicate that the latest models—such as Gemini 3 Pro, Gemini 2.5 Pro, GPT-5, o4-Mini, and Claude Sonnet 4—are now capable of effectively identifying underlying rules. In contrast, weaker models still tend to persist with outdated rules after they have become invalid or struggle to consistently apply a new rule once it has been discovered [**?**, **?**]. Comparing these findings with previous research, it is evident that as the capabilities of new models continue to advance, there is a corresponding and significant improvement in their cognitive flexibility.

De Langis et al. (2026) also investigated working memory and rule-switching capabilities. Using the Wisconsin Card Sorting Test (WCST) human data from Barceló et al. (1997) as a reference, the researchers estimated a human accuracy rate of approximately 0.77 based on reported mean error counts. In contrast, the models they tested—Gemma2-9B, Llama3.1-8B, and Qwen2-7B—achieved accuracy rates ranging only between 0.12 and 0.52 (De Langis et al., 2026). However, it is worth noting that these three models do not belong to the current top-tier of mainstream Large Language Models (LLMs); consequently, these test results may not be comparable to the performance of more powerful state-of-the-art models.

Overall, the studies by Haznitrama et al. (2026) and De Langis et al. (2026) both utilize extended versions of the Wisconsin Card Sorting Test (WCST) to reveal an uneven distribution of foundational cognitive abilities across various models. We contend that within these comprehensive cognitive assessments, the significance of rule-switching tasks lies in their ability to test whether a model

can transcend basic memory and information comprehension to reach a level where it can actively utilize information to adjust its behavior. Synthesizing the preceding analysis, it appears that Large Language Models (LLMs) may have already achieved a high degree of proficiency in memorizing and understanding rule changes.

information, but they may not yet be able to stably modify actual strategies during continuous tasks \cite{De Langis et al., 2026; Haznitrama et al., 2026;

H. Li et al., 2025}. However, this ability is precisely the critical criterion for determining whether executive control capabilities have reached maturity.

6.4 功能固着、心理定势与洞察式问题解决

Functional Fixedness and Insight Problem Solving

In psychology, functional fixedness and insight problem solving both refer to a specific category of cognitive challenges: the tendency for individuals to be influenced by past experiences during problem solving. This often results in being constrained by the first rules or solutions that come to mind, making it difficult to think outside of conventional frameworks. Functional fixedness was originally used to explain why people struggle to perceive familiar objects as new tools. For example, when seeing a box, individuals may only think of its function for containing items and find it difficult to realize it could also serve as a support stand [**?**, **?**].

Mental sets, on the other hand, emphasize the impact of prior successful experiences on subsequent problem solving. If an individual has repeatedly used a specific method to solve problems, they may continue to apply that old method even when a simpler alternative becomes available [**?**, **?**].

Insight problem solving focuses on whether an individual can suddenly restructure their understanding of a problem to discover previously ignored conditions or new perspectives. Wallas (1926) described the insight process as consisting of several stages: preparation, incubation, illumination, and verification [**?**, **?**]. Later research on linguistic insight problems frequently utilized riddles to test whether individuals could escape the misleading nature of a problem' s surface narrative [**?**, **?**]. Consequently, research in this field primarily investigates whether Large Language Models (LLMs) are easily led astray by familiar expressions or common solution patterns.

1. Direct Transfer Studies of Classic Insight and Mental Set Paradigms

Orrù et al. (2023) were among the first to use linguistic insight problems to test GPT-3.

5. Their study employed linguistic insight tasks from Ansburg and Dominowski (2000) and compared the performance of GPT-3.5 with the

small-sample human performance reported in the original article. These problems are typically short and superficially disguised as arithmetic or common-sense questions. The true difficulty lies in identifying hidden conditions within the prompt. For instance, if asked "how many plums are on a pear tree," the correct reasoning is to realize that pear trees do not bear plums; if asked "how much dirt is in a hole," the key is that a hole, by definition, contains no dirt. The task included 15 practice problems followed by 15 transfer problems, the latter of which tested whether subjects could apply the problem-solving experience gained from the first set to a new set of problems [**?**, **?**]. The results showed that the model correctly answered 12 out of 30 questions, with an overall score close to the human average in the original study. However, the model' s response patterns did not fully align with those of humans, particularly in the transfer tasks. Furthermore, since these riddles are common on the internet, the model may have encountered similar answers in its training data [**?**, **?**]. Thus, while early versions of ChatGPT exhibited a level of insight approaching that of humans, their internal reasoning processes remained distinct from human cognition.

Haq et al. (2025) adopted a classic mathematical mental set paradigm from previous literature [**?**, **?**]. In this paradigm, some problems require multi-step calculations, while others can be solved using simpler shortcuts. The researchers aimed to investigate whether a model, after becoming accustomed to multi-step calculations in initial trials, would continue to apply complex steps to subsequent problems that could be simplified.

The researchers measured the performance of Llama-3.1-8B-Instruct, Llama-3.1-70B-Instruct, and GPT-4o, while also comparing changes in performance when using Chain-of-Thought (CoT) prompting. The results indicated that GPT-4o performed best overall. While the introduction of CoT generally improved accuracy across all three models, it also led them to apply complex calculation steps to problems where shortcuts were available. The researchers also designed conditions where participants solved complex problems before shortcut problems, and vice versa. The results showed that the former condition improved accuracy but did not significantly reduce the number of solution steps [**?**, **?**]. It appears that current LLMs still exhibit significant mental sets in mathematical tasks, which are difficult to mitigate even with CoT techniques.

2. Mental Set Research in Adapted and Extended Paradigms

Alavi Naeini et al. (2023) observed significant mental sets in an adapted task for divergent thinking. The study introduced the Only Connect Wall (OCW) dataset, derived from the "Connecting Wall" segment of the British television show *Only Connect*. This task is inherently similar to the Remote Associates Test (RAT), a classic paradigm for measuring divergent thinking. The problems include numerous intentionally designed "red herrings" —clues that appear to be correct answers but actually lead the solver toward incorrect groupings. The

researchers explicitly treated these designs as a method to induce mental sets.

The results showed that all models performed far below the level of humans on the show. While GPT-4 was the top performer, it still lagged significantly behind human experts [**?**, **?**]. Crucially, when the authors retested the models using the OCW-Randomized and OCW-WordNet datasets—which weaken or remove the interference of red herrings—model performance improved significantly. This demonstrates that the models'errors were indeed caused by misleading clues rather than a lack of familiarity with the cultural common sense required by the questions.

3. Practical Tasks Based on the Mental Set Framework

Shidara et al. (2026) examined mental sets within the context of clinical problem solving. The researchers designed "mARC" clinical problems that often include familiar clues intended to prompt a quick association with a common diagnosis or treatment plan. However, each problem also includes a critical condition that renders the common solution inapplicable in that specific context. If a solver follows the familiar clues, they are likely to fall into the trap of an incorrect option; if they notice the critical condition, they must re-evaluate the situation and, upon realizing the information is insufficient, choose to gather more data first.

The results indicated that the new generation of reasoning models performed better on these tasks. The performance of Claude 4.1 Opus, Gemini 2.5 Pro, GPT-5.1, and Grok-4-Fast-Reasoning showed no significant difference from the average of five physicians. Among them, Claude 4.1 Opus was the strongest, with an average accuracy of approximately 0.75, compared to the human physician average of approximately 0.66 [**?**, **?**]. Case analysis by the researchers also showed that stronger models were more likely to choose to continue collecting information when evidence was insufficient, rather than rushing to provide a diagnosis or treatment [**?**, **?**]. This suggests that the reasoning versions of the latest mainstream models demonstrate improved strategy adjustment capabilities in medical tasks. This clearly illustrates the rapid evolution of LLMs and underscores the necessity of replicating previous studies using the most recent models.

6.5 判断与决策偏差

In psychology, the study of judgment and decision-making biases primarily focuses on the systematic and regular errors individuals commit when faced with uncertain information, risky choices, or complex evaluations. Tversky and Kahneman (1974) proposed that people frequently rely on rapid judgmental cues, such as anchoring, representativeness, and availability; while these heuristics can enhance judgmental efficiency, they may also lead to stable biases [**?**, **?**]. Subsequently, through Prospect Theory, they further demonstrated that the framing of an identical outcome—whether presented as a gain or a loss—can

significantly alter an individual' s risk preferences [**?**, **?**].

In the classical paradigm, researchers have examined a wide range of cognitive phenomena beyond the anchoring effect, the representativeness heuristic, the availability heuristic, and framing effects. Notable examples include the Wason four-card selection task (Wason, 1960), which is used to observe confirmation bias, as well as numerous other newly discovered biases that have emerged in subsequent literature; these are not detailed further here.

1. Direct Transfer Studies of Classical Judgment and Decision-Making Bias Paradigms

Suri et al. (2024) investigated five categories of classical judgment biases—anchoring, representativeness, availability, framing effects, and the endowment effect—using GPT-3.

5. To mitigate the risk of the model simply recalling original test answers from its training data, the researchers designed new problems with similar underlying structures rather than using verbatim copies of original tasks. They further validated these new items with 220 human participants to ensure they successfully induced the intended biases in the same direction as the original paradigms (Suri et al., 2024).

The results demonstrated that GPT-3.5 exhibited judgment biases identical to those of humans across all tasks. This suggests that classical bias paradigms are effective at capturing systematic error tendencies in the outputs of Large Language Models (LLMs) (Suri et al., 2024). However, as this study focused primarily on GPT-3.5 with only limited supplementary testing on GPT-4, the findings primarily characterize the performance of early mainstream models within classical heuristic tasks.

Nguyen (2024) provides further standardized evidence regarding the anchoring effect. The researcher directly employed the classic two-stage numerical anchoring paradigm design (Jacowitz & Kahneman, 1995). In this paradigm, the model is first presented with either a high-value or low-value anchor point, after which it is asked to provide a numerical estimate. The objective is to observe whether the model' s subsequent estimate is influenced by the previously provided numerical value. The researcher designed the study to···

The primary research context focuses on financial forecasting, which carries significant practical implications as users increasingly rely on Large Language Models (LLMs) to make judgments based on historical prices, growth rates, or market trends. Findings indicate that GPT-4, Claude 2.0, Gemini Pro, and GPT-3.5 are significantly influenced by anchoring effects during numerical prediction tasks. While the application of specific prompting techniques can mitigate this bias to some extent, they fail to eliminate the effect consistently \cite{J.

K. Nguyen, 2024}. This study provides robust evidence that LLMs exhibit anchoring effects in financial numerical forecasting, serving as a critical cautionary note for applications such as AI-driven stock trading.

O' Leary (2024) applied the Wason discovery task (the 2-4-6 problem) to investigate confirmation bias in large language models. The researchers conducted tests on Claude 3.5, Gemini, and GPT-4o; however, due to the limited number of test items, the study is characterized as a preliminary exploratory analysis.

The results indicate that Claude 3.5 and Gemini initially interpreted the rules narrowly. When prompted to explore other potential rules, these models primarily modified their original conjectures rather than actively seeking counterexamples that might falsify their initial hypotheses. In contrast, ChatGPT-4o demonstrated superior performance by proposing broader rules \cite{O' Leary, 2025b}. These findings suggest that in small-sample testing, certain Large Language Models (LLMs) exhibit a tendency to confirm their first-generated hypotheses during open-ended tasks, rather than flexibly exploring a wider range of possibilities.

Macmillan-Scott and Musolesi (2024) evaluated GPT-3.5, GPT-4, Bard, Claude 2, and Llama 2 by requiring these models to complete 12 classic judgment tasks from cognitive psychology, including the Wason selection task, the Linda problem, and the Monty Hall problem. The researchers did not merely assess the correctness of the models' answers; they also focused on whether the error patterns exhibited by the models aligned with the common cognitive bias patterns observed in humans during these classic experiments.

The results indicate that GPT-4 exhibited the best overall performance, with 69.2% of its responses being both correct and supported by sound reasoning. More importantly, researchers discovered that the majority of the model' s errors were unrelated to typical human cognitive biases. Instead, the model more frequently exhibited computational errors, logical inconsistencies, discrepancies between explanations and final selections, or instability in answers when presented with the same question repeatedly [**?**, **?**]. This study serves as a critical reminder that when a model's underlying reasoning and stability are not yet fully mature, its failures in judgment tasks may stem primarily from fundamental errors—such as calculation mistakes and logical lapses—rather than higher-level cognitive biases. Only by first isolating these foundational errors can we accurately determine whether a model truly manifests judgment biases analogous to those found in humans.

(2) Adaptation and Extension of Integrated Research on Judgment and Decision-Making Biases. Chen et al. (2025), Huang et al. (2026), Li Hao et al. (2025), Knipper et al. (2025), and Shaikh et al. (2024) have all extended the study of judgment bias into more complex multi-task evaluation frameworks.

Chen et al. (2025) covered 18 categories of cognitive biases and compared model performance when facing these biases in both classical and operations manage-

ment scenarios. Their results indicate that while GPT-4 aligns more closely with normative answers in problems involving explicit computational methods, it continues to exhibit significant biases in tasks related to preferences, risk assessment, and logical testing. Huang et al. (2026), on the other hand, focused their analysis on anchoring···

fixed effects, using synthetic data to compare different sources of anchors. Their findings demonstrate that anchoring effects are not limited to real-world forecasting tasks but also consistently appear in more controlled experimental materials. Li Hao et al. (2025) provided evidence from Chinese text, discovering that the proportion and presentation order of positive and negative information in realistic materials—such as genetic testing and autonomous driving—can influence model judgment and induce significant framing effects. From this body of research, we can further confirm that the judgment biases of Large Language Models (LLMs) are not stable; rather, they fluctuate according to task type, material context, prompt details, and model versions \cite{Y. Chen et al., 2025;

L. Hao et al., 2025;

Y. Huang et al., 2025}.

Comprehensive evaluations by Knipper et al. (2025) and Shaikh et al. (2024) further demonstrate that while expanding task coverage helps reveal this inherent complexity, it also introduces new challenges. Specifically, different bias tasks vary significantly in nature; if they are simply aggregated into a single composite score, it becomes difficult to discern whether a model' s performance is hindered by anchoring effects, framing effects, or fundamental limitations in logical reasoning (Knipper et al., 2025; Shaikh et al., 2024). This suggests that the measurement of cognitive biases in Large Language Models (LLMs) cannot rely on the reductive scoring methods typical of traditional engineering benchmarks. Instead, such evaluations must be rigorously categorized and analyzed based on psychological constructs to ensure the resulting conclusions are meaningful.

3. Practical Tasks Based on the Judgment and Decision-Making (JDM) Bias Framework

This area of research focuses on identifying the specific cognitive biases that emerge in scenarios where users are likely to employ Large Language Models (LLMs) in real-world applications. Rather than focusing on abstract logical puzzles, these studies examine how the interaction between humans and AI influences decision-making quality and psychological tendencies during practical tasks.

Lou and Sun (2026) investigated the anchoring effect in realistic predictive numerical judgments, focusing on whether models are influenced by pre-existing numbers when performing practical estimation tasks. Rather than strictly following the classic two-stage numerical anchoring paradigm, the researchers utilized 62 real-world prediction problems covering topics such as weather, stock

prices, flight schedules, and social media engagement metrics. The study evaluated a wide range of models, including GPT-4o, GPT-4, GPT-3.5 Turbo, GPT-o3, DeepSeek-R1-Qwen-32B, Gemini-2.5-flash, Gemini-2.5-flash-lite, Claude 3 Haiku, and Claude 3.5 Haiku.

For each problem, the models were provided with different types of prompts: some contained relevant facts (such as the previous week' s highest or lowest price), some included expert predictions, and others provided weakly related background information. The results indicated that the intensity of the influence varied across different prompt types. While relevant factual prompts—such as historical highs and lows—did induce anchoring in subsequent estimates, expert opinion prompts exerted the strongest and most consistent influence overall [**?**].

The researchers also tested several intervention methods, including Chain-of-Thought (CoT) prompting and explicit instructions to ignore potential anchors; however, these approaches showed limited effectiveness. The only relatively effective mitigation strategy was the simultaneous presentation of both high and low anchors, which prevented the models from anchoring to a single piece of information.

O' Leary (2025) applied the concept of anchoring to patent scoring tasks. In this study, ChatGPT-4, Claude 2.1, and Gemini Pro were tasked with providing multi-dimensional scores for patent abstracts. After the initial assessment, expert scores were provided to the models as anchors. The results demonstrated that the models frequently adopted the expert values directly for the target dimensions; furthermore, this influence sometimes spilled over, causing adjustments in other unrelated scoring dimensions (O' Leary, 2025a). This research closely mirrors real-world usage scenarios, and its conclusions serve as a significant cautionary note for the integration of AI in professional evaluation tasks.

Choi et al. (2025) discuss the false consensus effect within the context of social judgment. The false consensus effect refers to the tendency of individuals to overestimate the extent to which others agree with their own perspectives. Their findings indicate that when models such as GPT-4, Claude 3 Opus, LLaMA 2 70B, and Mixtral 8x7B are placed in specific social choice scenarios, they tend to overestimate the prevalence of their assigned positions. This behavior is consistent with the false consensus effect observed in humans (Choi et al., 2025).

However, it should be noted that this study did not allow the models to choose a position freely before estimating the choices of others; instead, the models were assigned a specific stance beforehand. Consequently, the study provides only indirect evidence that these models possess an egocentric attribution mechanism similar to that of humans.

Hwang et al. (2026) and Lior et al. (2026) both investigate whether Large Language Models (LLMs) exhibit framing effects during evaluation tasks. Specifically, Hwang et al. (2026) examined how model performance changes when the object of evaluation remains constant, but the framing of the prompt shifts from

an affirmative to a negative phrasing. Their results demonstrate that several mainstream models exhibit significant inconsistencies across both conditions when performing tasks such as judging the truthfulness of arguments, safety detection, toxicity detection, and grammatical assessment [**?**]. These findings indicate that LLMs, when acting as evaluators, are substantially influenced by the specific wording of a question. Given the increasing prevalence of using LLMs as automated scoring tools (LLM-as-a-judge) in current research, the implications of this study serve as a critical warning regarding the reliability of such evaluation frameworks.

Lior et al. (2026) investigated whether Large Language Models (LLMs) are susceptible to framing effects during emotion recognition tasks. In their study, they appended positive or negative supplementary descriptions to authentic Amazon user reviews and subsequently tasked 11 different LLMs, alongside human annotators, with assessing the emotional shifts. The results demonstrated that both the models and human participants were significantly influenced by these supplementary descriptions, with the impact of positive framing being particularly pronounced (Lior et al., 2025). Both studies conclude that when LLMs are utilized for evaluative tasks, their outputs are markedly affected by the phrasing of the prompts and the provided contextual framing.

6.6 问题解决总体评价

Overall, existing research indicates that Large Language Models (LLMs) have demonstrated certain capabilities in the domains of problem-solving and executive control, though these abilities are unevenly distributed. In tasks involving planning and sub-goal decomposition, current evidence remains limited; while models can handle simple path relationships, they struggle to stably maintain complex state spaces or re-plan in response to dynamic changes [**?**, **?**]. Regarding rule switching and cognitive flexibility, stronger models have shown the ability to discover rules and execute a degree of rule switching. However, in continuous tasks, they remain prone to issues such as perseveration on old rules, unstable strategy execution, and performance degradation when processing visual materials \cite{De Langis et al., 2026; Goto et al., 2025; Haznitrama et al., 2026;

H. Li et al., 2025;

J.

K. Nguyen, 2024}.

In the context of functional fixedness, mental sets, and insight-based problem solving, LLMs occasionally provide insightful solutions comparable to those of humans. Furthermore, the latest reasoning models have demonstrated improved strategy adjustment capabilities in practical tasks. Nevertheless, these models remain susceptible to the influence of familiar cues, default solutions, and misleading information [**?**, **?**, **?**, **?**, **?**]. Regarding judgment and decision-making

biases, the body of evidence is most substantial.

Models exhibit systematic tendencies in tasks involving the anchoring effect, framing effect, confirmation bias, and social judgment bias. These biases vary significantly depending on the task context, prompting methods, material formats, and model versions \cite{Y. Chen et al., 2025; Choi et al., 2025;

L. Hao et al., 2025;

Y. Huang et al., 2025; Hwang et al., 2026; Knipper et al., 2025; Lior et al., 2025; Lou & Sun, 2026; Macmillan-Scott & Musolesi, 2024;

J.

K. Nguyen, 2024}.

Nguyen, 2024; O' Leary, 2025b, 2025a; Shaikh et al., 2024; Suri et al., 2024)。

Synthesizing the aforementioned research, we contend that the problem-solving capacity of Large Language Models (LLMs) currently resembles a form of reasoning grounded in linguistic representation, rather than a fully developed, stable executive control capability. Future evaluations of LLM problem-solving proficiency should move beyond static performance metrics to investigate whether a model can maintain goals across continuous tasks, monitor its own errors, proactively revise strategies, and preserve cognitive flexibility in the presence of distracting information. These dimensions may prove to be the critical factors in distinguishing mere "test-taking" ability from genuine problem-solving competence.

7.1 研究范围界定

(1) Conceptual Definition

Theory of Mind (ToM) refers to the ability of an individual to understand unobservable mental states—such as beliefs, knowledge, intentions, desires, emotions, and perceptions—and to use this understanding to explain or predict the behavior of others. Premack and Woodruff (1978) first introduced this concept, centering it on the core question of "whether an individual can understand that others possess mental states" [**?**, **?**]. Subsequently, developmental psychology further operationalized this question, with the "false belief" tradition becoming the most influential framework. The "Unexpected Location" task by Wimmer and Perner (1983) and the "Sally-Anne" task used by Baron-Cohen et al. (1985) both require children to distinguish between the actual state of reality and the action expectations formed by others based on false beliefs. These tasks have since become essential measures of explicit Theory of Mind ability [**?**, **?**, **?**].

As research has evolved, the measurement of ToM has expanded from first-order belief reasoning to second-order and higher-order mental state attribution. Second-order false belief tasks require individuals to process recursive mental states, such as "Person A thinks that Person B thinks···" , which demands a

greater capacity to maintain and update multiple beliefs compared to first-order tasks [?, ?]. Simultaneously, ToM research has gradually moved beyond structured questions—such as "whether someone knows something" or "where someone will look for an object" —to more complex social-pragmatic understanding. For instance, tasks involving the recognition of social faux pas, understanding of irony and sarcasm, white lies, indirect requests, implied intentions, character emotions, and social norm judgments require subjects to integrate others' knowledge states, communicative intentions, and emotional reactions within specific contexts [?, ?, ?].

Based on the psychological framework described above, we categorized our search into three sub-fields: belief reasoning, perspective-taking, and social cognition. The first two categories tend to involve more fixed structural paradigms at the cognitive level, while the latter encompasses a broader range of mental states within various social contexts.

(2) Introduction to Classic Theory of Mind Paradigms

Because the paradigms used in Theory of Mind are more complex than those in other fields, this section provides a preliminary introduction to the classic psychological paradigms discussed in the following sections.

Perspective-taking tasks focus on whether an individual can understand what information others have access to. The visual perception inference tasks developed by Masangkay et al. (1974) provided an early foundation for this category of research [?, ?].

In contrast, belief reasoning tasks focus on whether an individual understands what beliefs others form based on the information available to them. False belief tasks further require subjects to understand that others will form incorrect beliefs due to incomplete information when their beliefs conflict with reality. The underlying design of the Sally-Anne type false belief task is that when a protagonist does not witness an object being moved, the subject must predict the protagonist's searching behavior based on that protagonist's false belief [?, ?]. The "Smarties" and "Unexpected Contents" tasks require subjects to distinguish between the actual contents of a container and the false beliefs others form based on its external packaging [?, ?]. Higher-order false belief tasks require subjects to recursively attribute multiple layers of mental states, such as "A thinks that B thinks an object is in a certain place" ; a representative task for this is the Imposing Memory Task (IMT) [?, ?].

Social cognition ToM paradigms emphasize complex interactive contexts. The Faux Pas Recognition Test requires subjects to identify socially awkward remarks, explain why they are inappropriate, and infer the speaker' s intentions as well as the feelings of the affected party [?, ?]. The Strange Stories task, Hinting Task, Irony/Sarcasm tasks, and Story Comprehension Test require subjects to understand non-literal language, indirect requests, sarcasm, white lies, threats, jokes, and implied intentions [?, ?, ?].

Together, these paradigms constitute the frame of reference for evaluating the performance of Large Language Models (LLMs) on Theory of Mind tasks in the subsequent sections.

7.2 过往综述

The studies included in this section are review articles previously conducted on the measurement of Theory of Mind in the field of Large Language Models (hereafter referred to as LLM-ToM). These works provide a valuable methodological framework for our reference.

(1) A Complex Research Landscape: Existing LLM-ToM literature has formed a complex research landscape, encompassing surface-level behavioral performance, internal engineering mechanisms, and theoretical articles discussing methodology and ethics.

Nguyen (2025) categorizes relevant research into three types: behavioral assessment, representational interpretation, and safety risks. Behavioral assessment focuses on model performance in tasks such as false belief, sarcasm understanding, higher-order mental state reasoning, and various comprehensive Theory of Mind benchmarks. Representational interpretation attempts to decode self and other belief states from the model' s internal representations. Safety risks focus on the potential privacy, deception, and anthropomorphism risks that may arise if models can infer user beliefs, preferences, or intentions \cite{H.

M. Nguyen & others, 2025}.

The systematic review by Sarıtaş et al. (2025) further demonstrates that current LLM-ToM evaluations are highly concentrated on text-based question-answering tasks using the GPT series. They point out that most studies employ multiple-choice, true/false, or open-ended short-answer formats, while only a few involve images, 2D/3D environments, or multi-agent interactions [**?**, **?**].

Marchetti et al. (2025) attempt to answer through their review whether the "understanding" exhibited by LLMs is merely an illusion. They emphasize that the ToM-like performance of LLMs likely stems from the superficial exploitation of linguistic patterns in social contexts; furthermore, researchers often overinterpret a model' s score performance as evidence of an actual underlying psychological mechanism.

(2) Controversies in LLM-ToM Research: Hu et al. (2025) argue for a reexamination of existing ToM assessments for LLMs, emphasizing that a model' s generation of text matching a standard answer only indicates that it completed the corresponding response under those task conditions. It does not yet prove that the model represents the beliefs of others in the same manner as human children or adults [**?**, **?**]. From the perspective of general users, Wang et al. (2025) point out that many benchmarks remain dominated by a static, third-person tone. These benchmarks rarely

consider how real users in interactions require models to identify misunderstandings and revise their understanding based on multi-turn dialogue history. There remains a gap between a model' s score on standard paradigm questions and its social understanding in real-world human-computer interactions [**?**, **?**].

Yin et al. (2025) further criticize the conflation of the concepts of "Mental Model" and "Theory of Mind" in LLM-ToM research. A Mental Model is a concept at the representational level, focusing on the internal representation an individual builds in their own mind regarding an object, situation, system, or another' s cognition [**?**, **?**]. In contrast, ToM is a concept at the level of social cognitive ability, focusing on whether an individual can stably track the mental states of others across different contexts to predict behavior and revise their understanding of others based on feedback during interactions; it is comparatively more specific and dynamic. They point out that many so-called LLM-ToM studies actually measure surface-level behaviors that do not necessarily involve genuine mental state attribution [**?**, **?**].

Taken together, the greatest controversy within existing LLM-ToM research is: are we measuring whether the model can behave like a human, or are we measuring whether it possesses the underlying computational mechanisms to support such behavior?

(3) Prompting, Engineering Enhancements, and Evaluation Frameworks: Prompting methods and engineering enhancements are unavoidable factors when interpreting LLM-ToM results and are included here as supplementary literature. Chen et al. (2025) organized LLM-ToM research from the perspectives of evaluation and enhancement, noting that prompting methods such as Chain-of-Thought (CoT), Few-Shot, and Perspective-Taking Prompts, as well as certain external tools, can significantly alter a model' s score on ToM benchmarks [**?**, **?**]. Similarly, the aforementioned Sarıtaş et al. (2025) summarized prompting frameworks such as SimToM, TimeToM, PercepToM, and SymbCoT. The commonality among these methods is that they decompose the mental state reasoning process—which the model would otherwise need to complete autonomously—into several more explicit steps to improve task performance.

7.3 视角采择

(1) Direct Transfer Studies of Classical Perspective-Taking Paradigms

Compared to other subfields, LLM research that purely replicates classical visual perspective-taking tasks has not yet gained significant prominence. This may be because classical visual perspective-taking requires the perception of spatial positions, relies on visual visibility, and necessitates the analysis of body ori-

entation and occlusion relationships within a scene. However, most text-based LLMs can only process transcribed linguistic descriptions. Consequently, the so-called "perspective-taking" in many studies is often intertwined with the extraction of textual cues and spatial modeling, or is measured using autonomous agents; neither of these aligns with our definition of direct measurement.

(2) Adapted and Extended Perspective-Taking Research

Jung et al. (2024) attempted to distinguish between the processes of information acquisition and belief formation in LLMs by modifying classical belief tasks. This study decomposed Theory of Mind (ToM) reasoning into two stages: Perception Inference and Perception-to-Belief Inference. The former examines whether a model can determine what a character sees or hears, while the latter examines whether the model can infer a character' s beliefs based on their perceptual state.

After testing models such as GPT-3.5 Turbo, GPT-4 Turbo, GPT-4o, Claude 3 Haiku, Claude 3 Sonnet, and Gemini 1.0 Pro, the results indicated that models generally identify information perceived by characters effectively. However, they struggle to consistently infer a character' s beliefs from that perceived information. Specifically, they are highly susceptible to interference from factual information about reality that the target character has not perceived [**?**]. This study effectively distinguishes perspective-taking from belief reasoning, finding that while the models' perspective-taking abilities are sufficient, they frequently falter at the step of transitioning from "information in another' s eyes" to "beliefs in another' s mind."

In other words, within the framework of Theory of Mind capabilities, these models exhibit sufficient perspective tracking but insufficient belief updating.

(3) Practical Tasks Based on Perspective-Taking Frameworks

Verma et al. (2024) utilized a Perceived Behavior Recognition task within a human-computer interaction scenario. The setup involved providing the model with a robot' s action plan or behavior description and then asking the model whether a human observer would find the robot' s behavior intelligible, predictable, goal-oriented, or intended to hide a true objective. Essentially, the model is required to adopt the position of a human observer to judge how certain robotic behaviors would be interpreted by humans.

The researchers tested GPT-4 and GPT-3.5-turbo alongside a concurrent study involving human subjects. The testing was conducted in two stages. First, under standard conditions, the models were asked to directly judge which category a robot's behavior belonged to or whether it matched specific behavioral characteristics based on the provided descriptions. Under these conditions, both GPT-4 and GPT-3.5-turbo performed at levels highly similar to humans [**?**]. The second

stage introduced perturbation conditions, where researchers added information that theoretically should not change the answer—such as task-irrelevant context, belief descriptions conflicting with the original situation, or requiring the model to maintain its judgment after it had already been given. The results showed that all models became significantly unstable under these perturbations, suggesting that their high scores in standard conditions do not necessarily stem from robust second-order mental state judgments [**?**].

This research provides evidence from practical scenarios for the LLM-ToM literature, demonstrating that the stability of models in these reality-oriented perspective-taking tasks is easily compromised by situational perturbations.

7.4 信念推理

(1) Direct Transfer Studies of Classical False Belief and Higher-Order Belief Paradigms. False belief and higher-order belief reasoning are the most central sources of evidence in LLM-ToM research, aligning most closely with classical psychological traditions.

Street et al. (2024) investigated the 2nd- to 6th-order ToM attribution capabilities of LLMs using an Imposed Memory Task. This study covered everything from classical second-order false beliefs to higher-order recursive beliefs, comparing LLM performance with that of 29,259 British adults. To rule out the possibility that models simply excel at memorizing complex nested sentence structures, the researchers also implemented corresponding 2nd- to 6th-order factual statement questions. These required the models to determine the validity of multi-layered factual relationships (e.g., "The red box is next to the blue box, the blue box is behind the green box, and the green box is under the table. Is the statement 'the red box is under the table' true or false?") without involving mental state terminology.

The researchers tested LaMDA, PaLM, Flan-PaLM, GPT-3.5 Turbo Instruct, and GPT-4. Overall, GPT-4 reached the average adult level across 2nd- to 6th-order tasks and even exceeded average adult performance on 6th-order ToM tasks. Specifically, human accuracy was approximately 97.5% for factual questions and 90.4% for ToM questions; GPT-4 achieved approximately 94.3% on factual questions and 88.6% on ToM questions; and Flan-PaLM reached approximately 93.8% on factual questions and 84.3% on ToM questions [**?**]. These data indicate that ToM tasks incorporating a belief-understanding dimension do indeed introduce additional difficulty. However, GPT-4' s near-adult performance under these more challenging conditions suggests it possesses the capacity to handle nested higher-order mental states in text-based belief reasoning tasks [**?**]. With its comprehensive measurements and large-scale human baseline for reference, this study provides highly robust evidence within the field of belief reasoning.

(2) Adapted and Extended Belief Reasoning Research. In addition to the direct transfer of classical paradigms, researchers have expanded false belief

logic into larger-scale, more templated, and multilingual tasks.

Suzgun et al. (2025) constructed the KaBLE benchmark, consisting of 13 categories of epistemic tasks and approximately 13,000 questions, to test whether 24 mainstream LLMs could distinguish between belief, knowledge, and fact. The researchers directly examined whether models could consistently judge the differences between "someone believes X," "someone knows X," and "X is a fact." The results showed that while models performed relatively well on third-person narrative false beliefs, they were significantly unstable in first-person narrative conditions. For example, GPT-4o' s accuracy dropped from 98.2% to 64.4% between the two conditions, while DeepSeek R1 plummeted from over 90% to 14.4%—a common flaw in the new generation of models. This suggests that when encountering false beliefs (believing something that is factually false), models are still easily influenced by the factual dimension, tending to correct the false proposition rather than recognizing that the speaker genuinely holds that erroneous belief [**?**].

Gandhi et al. (2023) proposed BigToM, which further subdivides the logic of false belief tasks into three types of belief judgments. The first is Forward Belief inference ("What will the character believe after seeing this information?"). The second is Forward Action inference ("How will the character act given these beliefs and goals?"). The third is Backward Belief inference ("Given that the character performed a certain action, what might they believe?"). After establishing these three directions—from information to belief, from belief to action, and from action back to belief—the researchers generated a large number of narrative questions and compared the performance of five LLMs against 20 human subjects in a forced-choice format. The results indicated that GPT-4 was the model most closely resembling human reasoning patterns, followed by Claude [**?**]. In Forward Belief inference, GPT-4 and Claude approached human levels, while other models were more prone to failure. In Forward Action inference, only GPT-4 slightly outperformed humans under false belief conditions, while other models performed poorly. In the most difficult task, Backward Belief inference, human performance also declined, but all models performed significantly below the human average [**?**].

Similar to the previous study, by decomposing the reasoning process of belief tasks, this research found that belief reasoning in LLMs is not a monolithic capability but varies according to the direction of inference; furthermore, the models tested at the time could only approach, but not yet truly reach, human levels. Milička et al. (2024) adopted a more innovative design using persona prompting to have models simulate subjects of different ages, requiring the models to adjust their performance on false belief tasks to match their assigned persona. After testing GPT-3.5-turbo and GPT-4, they compared the models' performance with actual child development curves from developmental psychology research and meta-analyses [**?**, **?**]. The consensus in these developmental studies is that children generally fail around age 3, show significant improvement after age 4, and reach maturity by ages 5–6. The results showed that

GPT-4' s simulations closely mirrored real child development curves, although it occasionally produced excessively high accuracy when portraying a 3-year-old [?]. This study holds methodological significance, as persona prompting can indeed examine a model' s ability to simulate a target' s mental state to some extent. However, such testing remains an indirect measure, and results may be influenced by external factors such as the model' s knowledge base, thus the strength of the evidence remains limited.

Sadhu et al. (2024) constructed Multi-ToM based on a subset of ToMBench, covering English, Arabic, French, Hindi, Bengali, Russian, and Chinese. The tasks include false belief, implication, strange stories, social faux pas, and non-verbal communication [?]. Chan et al. (2025) also proposed XToM, extending comprehensive belief tasks such as ToMi, FANToM, and NegotiationToM to multilingual scenarios to examine whether LLMs can consistently perform mental state reasoning across different languages [?]. Additionally, Ünlütabak and Bal (2025) measured first- and second-order false belief tasks in both English and Turkish using GPT-3.5 and GPT-4. Their results showed that the models' strong performance on standard English tasks did not transfer to the Turkish context [?]. Collectively, these studies indicate that LLM performance under multilingual ToM assessments is jointly influenced by translation quality, the linguistic resources available during training, and cultural context [?, ?]. This suggests that current mainstream LLMs, primarily trained on English corpora, still exhibit language dependency in belief reasoning.

7.5 广泛心理状态归因

1. Direct Transfer Studies of Classical Social Cognitive Paradigms

Holl-Etten et al. (2025) utilized three classical social cognitive paradigms—the Faux Pas Recognition Test, the Social Stories Questionnaire, and the Story Comprehension Task—to evaluate the performance of GPT-3.5 Turbo and GPT-4 in processing complex social information under both English and German language conditions.

The performance of Large Language Models (LLMs) on social short stories has been a subject of rigorous investigation. Researchers strictly followed original testing manuals for scoring, with evaluations conducted by two independent raters who maintained high inter-rater reliability. Results indicate that GPT-4 demonstrates strong overall performance in interpreting complex social narratives. It is capable of not only identifying whether a character has said something inappropriate but also frequently providing correct explanations for why the statement was unsuitable, the speaker' s underlying motivations, and the potential emotional impact on the offended party. In contrast, GPT-3.5 exhibited significantly more errors in these areas, showing particular weakness in tasks requiring the comprehension of social faux pas, implicit intentions, and non-verbal cues [?, ?]. By utilizing both English and German materials, the study

tested whether the models were merely familiar with original English tasks or overly dependent on English training data; the findings confirmed that GPT-4 performed robustly across both languages [**?**, **?**].

Furthermore, Holl-Etten et al. (2025) observed that while GPT-4' s answers often received high scores, the model frequently employed hedging language and uncertainty markers such as "possibly," "maybe," or "probably." This observation is critical for practical applications, as it suggests that a model' s ability to provide a correct answer for scoring purposes does not necessarily guarantee it can deliver social insights in a stable, clear, or user-appropriate manner.

In comparison, the study by Strachan et al. (2024) featured a larger experimental scale, although it utilized older model versions. The researchers employed five categories of classic Theory of Mind (ToM) tasks to compare the performance of GPT-4, GPT-3.5, and LLaMA2-70B-Chat against a human sample of 1,907 participants. These five categories included four social cognitive paradigms: irony comprehension, hint tasks, strange stories, and social faux pas recognition. To mitigate the possibility of the models relying on direct memorization of original test items, the authors generated new materials for several of the publicly available tasks.

The results indicate that GPT-4 exhibits the strongest overall performance, reaching or exceeding human-level proficiency in tasks involving irony comprehension, indirect hints, and strange stories. GPT-3.5 followed as the second-best performer, while LLaMA2-70B-Chat showed comparatively weaker results [**?**, **?**]. Notably, GPT-4 initially performed below human levels in social faux pas tasks; however, subsequent controlled experiments revealed that this failure was partially attributable to the specific phrasing of the questions and the model' s inherently conservative default response strategies [**?**, **?**]. This highlights that Large Language Model (LLM) judgments regarding social norms are highly sensitive to the framing and elicitation methods used in the prompts.

2.2 Adapted and Expanded Research on Broad Mental State Attribution

Jones et al. (2024) introduced a comprehensive benchmark named EPITOME, which compares human performance against five models from the GPT-3 series, utilizing GPT-3 *text-davinci-002* for their primary analysis. This integrated benchmark comprises six categories of Theory of Mind (ToM) tests, encompassing second-order false belief and higher-order belief reasoning. It also includes social cognitive tasks such as short story comprehension, "Strange Stories," and the interpretation of indirect requests. To ensure a rigorous comparison, the authors identified human baseline data from previous studies for each task category.

The results indicate that the performance of the GPT-3 series models is inconsistent across different ToM tasks. While the models approach human-level performance in certain story comprehension and lower-order belief tasks—and

even surpass humans in short story reasoning—they lag significantly behind in higher-order belief reasoning, the understanding of indirect requests, and social semantic judgments [**?**]. However, as previously noted, we have observed that···

Subsequently, GPT-4 achieved significant breakthroughs in high-order belief reasoning and various social cognitive tasks [**?**, **?**, **?**, **?**].

Tong et al. (2026) proposed a comprehensive Theory of Mind (ToM) assessment framework called CogToM, which comprises 46 task paradigms categorized into 7 core abilities and 36 sub-abilities. While these tasks are derived from established research in the field of psychology, our study focuses specifically on the more classic paradigms within this set. After evaluating 22 representative Large Language Models (LLMs), the researchers observed highly uneven performance across different models. On one hand, the rapid evolution of these models is evident: early models, such as the Llama-2 series, exhibited low average scores, whereas more recent frontier models like Qwen3-Max and GPT-5.1 have achieved scores exceeding 80% (Tong et al., 2026).

On the other hand, model progress has been concentrated within specific task types. For instance, performance on traditional false-belief tasks has nearly reached saturation, with many powerful models achieving high scores. Models also demonstrate a strong proficiency in utilizing textual cues to answer tasks involving the judgment of emotions, intentions, and non-verbal communication (Tong et al., 2026). In contrast, tasks involving spatial perspective-taking and inferring beliefs from perception remain significant weaknesses. This latter finding aligns with the results reported by Jung et al. (2024), suggesting that these areas represent genuine bottlenecks in current model evolution.

Li et al. (2025) presents a comprehensive study that evaluates Large Language Models (LLMs) through a psychometric framework. The researchers assessed five distinct psychological dimensions: personality, values, emotions, Theory of Mind (ToM), and motivation. The study tested nine mainstream LLMs, including GPT-3.5-Turbo, GPT-4, GLM4, Qwen-Turbo, Mistral-7B, Mixtral 8$×$22B, and Llama3-70B.

Within the ToM sub-module, Li et al. (2025) utilized three specific categories of tasks: False Belief tasks, Strange Stories tasks, and Imposing Memory tasks. The results indicated that while GPT-4 and Llama3-70B demonstrated overall superior performance, significant variations were observed across different tasks [**?**].

It is noteworthy that researchers have introduced minor perturbations to these classic tests to examine the stability of the results. For instance, they investigated whether models would still select the same answer after changing the position of options in multiple-choice questions, or whether results remained consistent after replacing character names and genders. The results most significantly affected were multiple-choice Theory of Mind (ToM) tasks, particularly those involving false beliefs and imposed memory. Because both types of tasks require the model to select an answer from a given set of options, the position

of the options, their order, and surface-level paraphrasing ultimately directly influenced whether the model maintained a consistent judgment [**?**, **?**]. This study serves as a reminder that the interpretability of ToM scores can change significantly depending on whether the question format is prone to inducing positional bias or framing effects. Consequently, the stability of the scoring method determines whether the results are truly robust.

3. Practical Tasks Based on a Comprehensive ToM Framework

Salmani-Zarchi et al. (2025) investigated whether models could simulate Theory of Mind (ToM) performance corresponding to different levels of autism support requirements based on specific prompts. To achieve this, the authors constructed approximately 500 multiple-choice questions encompassing 11 distinct categories of ToM tasks. This framework provides comprehensive coverage of paradigms related to perspective-taking, belief reasoning, and social cognition. In this study, researchers tasked the models with adopting the personas of neurotypical individuals as well as individuals across varying levels of the autism spectrum to evaluate their behavioral consistency and simulation accuracy.

Individuals with subclinical autistic traits and those diagnosed with ASD Levels 1–3 according to the DSM-5 will be examined to determine whether the models exhibit distinguishable gradients in terms of question accuracy, response length, and the stylistic use of mental state terms. The models selected for testing include GPT-4o, Gemini-1.5-Pro, as well as the Llama, Gemma, Qwen, and DeepSeek series. The primary human benchmarks for this study are derived from diverse population data documented in existing clinical literature.

The results revealed that GPT-4o achieved the highest overall accuracy. Along with Gemini-1.5-Pro, it successfully simulated a progressive decline in performance across ASD Level 1, Level 2, and Level 3 profiles. These performance differences were primarily observed in more complex tasks, such as Strange Stories, Social Faux Pas, and Unexpected Content False Belief tasks [**?**, **?**]. In contrast, smaller open-source models struggled to significantly distinguish between the different diagnostic profiles under baseline conditions. Furthermore, the researchers discovered through fine-tuning that a model' s alignment strategy significantly influences simulation outcomes: models more heavily biased toward providing helpful and standardized responses are more likely to flatten the nuanced differences between various ASD levels [**?**, **?**].

This research holds significant practical implications for clinical training. For instance, it could be utilized in the future to develop virtual clients or neurodiversity-sensitive AI assessment materials. Such tools would assist trainees in practicing clinical interactions with patients who require varying levels of support. However, it is important to note that the ASD profiles simulated by the model are highly dependent on specific role-playing prompts and alignment methods. Consequently, practical applications will require the design of standardized prompting templates to ensure consistency and

reliability.

7.6 心理理论领域总体评价

Overall, research into Theory of Mind in Large Language Models (LLM-ToM) has evolved from early studies focused on simple false-belief tasks into a comprehensive field covering perspective-taking, belief reasoning, and broad social cognition. Current evidence indicates that newer LLMs can achieve high scores across many classic ToM tasks. Specifically, in areas such as text-based false belief, higher-order belief nesting, irony comprehension, understanding implications, and identifying social faux pas, some models have even approached or exceeded human baselines [**?**, **?**, **?**, **?**, **?**, **?**]. These findings suggest that LLMs possess robust capabilities for processing social-contextual text and can generate responses consistent with ToM logic in many standardized tasks.

However, from the perspective of psychological research, the high scores achieved by these models must be interpreted with caution. Most current tasks center on static text, third-person narratives, and standardized answers; consequently, they measure the stability of model performance within specific question formats rather than providing direct proof that models possess human-like mental state representation mechanisms. Models continue to exhibit significant instability, particularly in tasks involving inferring beliefs from perception, reverse belief inference, multilingual transfer, sensitivity to prompt perturbations, and real-world interaction scenarios. Furthermore, the ToM performance of LLMs remains subject to the collective influence of task type, linguistic environment, prompting methods, scoring formats, and model alignment techniques [**?**, **?**, **?**, **?**].

2024;

Y. Li et al., 2024; Sadhu et al., 2024; Salmani-Zarchi et al., 2025; Suzgun et al., 2025; Ünlütabak & Bal, 2025; Verma et al., 2024)。

We believe that, on the one hand, the rapid progress of models within classical Theory of Mind (ToM) paradigms should be acknowledged, as this provides new research tools for social cognitive measurement, educational training, and human-computer interaction. On the other hand, future research must advance in directions such as dynamic interaction, multimodal contexts, cross-cultural linguistics, robustness testing, and the interpretation of internal mechanisms. Only through such efforts can we more accurately determine whether Large Language Models (LLMs) are truly performing Theory of Mind tasks or are merely simulating ToM-like behavioral patterns within specific textual environments.

8 对过往研究的重新评估

Based on a systematic review, this study further conducted re-evaluations and minor perturbation tests on several representative studies. The reproducibility

tests covered 10 papers, while the minor perturbation tests covered 4 papers. All re-tests were performed using ChatGPT-5.4 without the "thinking mode" enabled. Except for the change in model version, all other experimental conditions followed the original settings as closely as possible. Due to the large volume of testing content, only brief conclusions are reported in the main text; detailed data are provided in the experimental reports in the Appendix.

In the domain of creative thinking, this study re-evaluated three papers. We tested the egg-drop experiment, divergent association tasks (DAT), alternative uses tasks (AUT), and creative connection tasks such as the Only Connect Wall. The results indicate that GPT-5.4 can generate a greater number of answers and cover a wider range of idea categories; its performance improves significantly when creativity is explicitly requested. For instance, in the egg-drop experiment, it covered all 10 categories of solutions; in the DAT, it reached a high human percentile under the "aware" condition; and in the AUT, originality scores for "creative prompts" were higher than those for "practical prompts." While GPT-4o in the original literature already demonstrated strong generative capabilities, it tended to cluster around conventional solutions. Compared to GPT-4o, GPT-5.4 exhibits a broader expansion range and a decreased proportion of fixed paths, suggesting a better ability to move beyond common answers. However, it still lacks the ability to consistently judge which answers are more creative, and it remains inferior to humans under the original misleading conditions of the Only Connect Wall. Therefore, updates in creative tasks are primarily reflected in the increased breadth of generated content, while understanding, filtering, and deep connection of creativity remain limited.

In the field of Theory of Mind (ToM), this study re-evaluated two papers. We tested the ToMi structured story task, the FANToM multi-party conversation task, and the BigToM social reasoning task, focusing on whether the model could judge what characters know, what they believe, and how they will act. The results show that GPT-5.4 performs exceptionally well in clearly structured story tasks, achieving an overall accuracy of 92.7% on ToMi, with particularly stable performance on false-belief questions. In the FANToM multi-party conversation task, its direct responses were stronger than those of previous models; in BigToM, it showed high performance in forward belief and forward action reasoning. While GPT-4o in the original literature possessed some capability in simple ToM tasks, it was weak in multi-party conversations and complex false-belief tasks, often requiring additional methods like PercepToM to extract character perspectives.

The direct responses of GPT-5.4 have already surpassed the "vanilla"condition of GPT-4o reported in the original literature, and on ToMi, they even exceeded the results of GPT-4o combined with PercepToM. However, performance remains unstable under the FANToM error-belief and BigToM reverse-belief reasoning conditions. We can conclude that the new model' s belief tracking in simple stories has significantly improved, but it still faces difficulties in multi-party conversation scenarios and in inferring beliefs backward from actions.

In memory metacognition tasks, we tested whether the model' s Judgments of Learning (JOL) for garden-path sentence pairs could truly predict memory difficulty. The results indicate that while GPT-5.4 can identify the presence or absence of semantic relationships between sentences, it fails to use this information to infer that a sentence will be easier to remember. Compared to the original literature, this continues a problem seen in earlier models: while human memorability judgments can predict subsequent memory performance, GPT-4o could not. Although GPT-5.4 demonstrates a better understanding of the relationships between materials, this does not yet indicate that it possesses human-like memory monitoring capabilities.

In analogical reasoning tasks, we tested whether the model could still grasp the same abstract relationships when encountering letter changes, symbol substitutions, matrix format variations, and story rewrites. The results show that GPT-5.4 has made significant progress in these tasks, particularly in letter analogies and numerical matrix tasks, where it is no longer easily distracted by surface-level format changes.

Compared to the original literature, where GPT-4' s performance often declined due to changes in gap positions, material formats, or answer sequences, GPT-5.4 is much more stable in these areas. This suggests that its analogical transfer capability is stronger than that of previous models. However, fluctuations still occur when a problem requires processing multiple symbolic rules simultaneously or when the order of options in story analogies is altered.

In problem-solving tasks, we tested the model's ability to learn through feedback, utilize probabilistic information, and adjust strategies after rule changes. The results show that GPT-5.4 performs strongly in these tasks: it can gradually avoid choices with poor long-term returns, select more advantageous options based on probabilities, and continue to correct its answers after rules change.

Compared to the original literature, this result largely continues the characteristics of earlier LLMs, where models often achieve high scores and even approach optimal strategies more closely than humans. The new development is that GPT-5.4' s tendency toward optimization is even more pronounced. Especially in risky decision-making, it behaves more like it is consistently pursuing maximum returns, rather than adjusting bets based on risk perception as humans do.

In anchoring effect tasks, we tested whether the model' s estimates would be influenced by cues provided earlier in the prompt. The results show that GPT-5.4 is relatively stable regarding irrelevant numbers but is still influenced by anchors containing semantic content. Compared to the original literature, the degree to which it is affected by semantic anchoring has decreased, suggesting that the new model has improved resistance to such cues. In terms of purely numerical anchoring, its performance is similar to the GPT-4o reported in the original literature, indicating no particularly significant change in this area. Overall, GPT-5.4 has not completely escaped the anchoring effect; it is simply less likely

to be misled by pure numbers.

In the minor perturbation tests, this study examined whether the model remained stable when faced with slight rewrites across four categories of tasks. We applied minor perturbations—including rewriting questions, adjusting sequences, changing material formats, or adding distracting information—to tasks involving verbal creativity, working memory and executive function, neurocognition, and high-order Theory of Mind. The results found that GPT-5.4 maintained high performance across most perturbed tasks: the average total score for verbal creativity perturbations was 81.97/100, the revised accuracy for working memory and executive function was 92.86%, neurocognitive perturbations reached 88.46%, and high-order Theory of Mind perturbations achieved 20/20. Errors were primarily concentrated in tasks such as operation span, n-back, and spatial working memory, all of which require the model to retain previous information while simultaneously updating its current state. While earlier studies often found LLMs to be sensitive to phrasing changes, order variations, and surface formats, these results indicate that GPT-5.4 is more stable against general linguistic rewrites and less prone to being misled by minor cues. At the same time, the model' s tendency to make errors during the continuous maintenance of internal states persists; the manifestation of this issue has shifted from sensitivity to perturbations toward instability in dynamic memory, positional updating, and distributional reasoning.

9 讨论与结论

Through a systematic review and re-evaluation, this study extensively delineates the boundaries of Large Language Models' (LLMs) capabilities within classical cognitive psychology paradigms. The superior performance of LLMs across various cognitive domains coexists with significant vulnerabilities, revealing a discrepancy between high scores at the behavioral performance level and the actual possession of underlying cognitive mechanisms.

The success of LLMs in standardized tests—such as creative association, deductive reasoning, and Theory of Mind—demonstrates that networks trained on massive text corpora can highly simulate human response patterns within specific contexts. However, this superficial similarity often masks structural deficiencies in complex, dynamic tasks. In tasks requiring continuous updates to working memory, resistance to mental sets, and the processing of multiple social interactions, models often struggle to maintain a coherent cognitive state. The anchoring and framing effects exhibited by models in cognitive bias tasks, along with the strategic fixation observed in practical scenarios such as medical or psychometric testing, further illustrate that their judgment processes rely heavily on superficial textual features.

Psychological theoretical frameworks provide valuable tools for evaluating complex intelligent systems, yet they also face challenges regarding conceptual validity. When classical psychological tasks are directly transferred to language

models, the measurement results are often confounded by language comprehension abilities and task-specific framing factors, due to the lack of embodied interaction and multimodal feedback within a real physical environment.

Future research in the psychology of artificial intelligence must develop dynamic, interactive measurement paradigms specifically designed for language models. Furthermore, evaluations must strictly distinguish between a model' s alignment with specific rhetorical logic and its actual computational processes.

From a practical application perspective, the instability exhibited by LLMs in perspective-taking and belief reasoning serves as a clear warning for high-sensitivity fields such as human-computer collaboration and clinical assistance. Models find it difficult to stably track a user' s deep intentions or revise previous erroneous hypotheses across multiple interactions. When handling tasks that require refined social cognition—such as clinical diagnosis, psychological counseling, or social

education—they may generate suggestions that appear plausible but lack substantive empathy and logical support. Researchers and developers must confront this imbalance in capabilities and establish rigorous boundary testing and safety monitoring mechanisms during application deployment.

In summary, while Large Language Models exhibit highly human-like cognitive behaviors, they have yet to show signs of possessing deep cognitive structures isomorphic to those of humans. The continuous scrutiny of LLM cognitive abilities will prompt the fields of cognitive science and artificial intelligence to further clarify the nature of intelligence, driving the development of more explanatory and robust artificial intelligence systems.

Abe, H., Ando, R., Morishita, T., Ozeki, K., Mineshima, K., & Okada,

M. (2024). Abductive reasoning with syllogistic forms in large language models. International Conference on Human and Artificial Rationalities, 3–17. Springer.

Abe, H., Ozeki, K., Ando, R., Morishita, T., Mineshima, K., & Okada,

M. (2026). Evaluation of Deontic Conditional Reasoning in Large Language Models: The Case of Wason' s Selection Task. Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), 588–601.

Ackerman,

C. (2026). Evidence for Limited Metacognition in LLMs. arXiv Preprint arXiv:2509.21545.

Anca, D., FLORESCU, A.-M., & REȘCEANU,

A. (2025). TEST DESIGN AND STRUCTURE FOR ASSESSING HUMANS AND LLMs' LINGUISTIC CREATIVITY. Annals of the University of Craiova. Series Philology. Linguistics, 47(1–Anderson,

J.

R. (1993). Problem solving and learning. American Psychologist, 48(1), 35–44. https://doi.org/10.1037/0003-066x.48.1.35 Ando, R., Ozeki, K., Morishita, T., Abe, H., Mineshima, K., & Okada,

M. (2024). Can Euler Diagrams Improve Syllogistic Reasoning in Large Language Models? International Conference on Theory and Application of Diagrams, 232–248. Springer.

Ansburg,

P. I., & Dominowski,

R.

I. (2000). Promoting insightful problem solving. The Journal of Creative Behavior, 34(1), 30–60.

Arora, V., Thabane, A., Parpia, S., Calic, G., & Bhandari,

M. (2025). Generative artificial intelligence models outperform students on divergent and convergent thinking assessments. Scientific Reports, 15(1), 36987.

Barbot, B., Besançon, M., & Lubart,

T. (2016). The generality-specificity of creativity:

Exploring the structure of creative potential with EPoC. Learning and Individual Differences, 52, 178–187. https://doi.org/10.1016/j.lindif.2016.06.005 Baron-Cohen, S., Leslie,

A. M., & Frith,

U. (1985). Does the autistic child have a "theory of mind" ? Cognition, 21(1), 37–46.

Baron-Cohen, S., O' riordan, M., Stone, V., Jones, R., & Plaisted,

K. (1999). Recognition of faux pas by normally developing children and children with Asperger syndrome or highfunctioning autism. Journal of Autism and Developmental Disorders, 29(5), 407–418.

Belém,

C. G., Kelly, M., Steyvers, M., Singh, S., & Smyth,

P. (2024). Perceptions of Linguistic Uncertainty by Language Models and Humans. In

Y. Al-Onaizan,

M. Bansal, & Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 8467–8502). Miami, Florida, USA: Association for Computational Linguistics.

https://doi.org/10.18653/v1/2024.emnlp-main.483 Bellemare-Pepin, A., Lespinasse, F., Thölke, P., Harel, Y., Mathewson, K., Olson,

J. A., ···Jerbi,

K. (2024). Divergent creativity in humans and large language models. arXiv Preprint arXiv:2405.13012.

Berg,

E.

A. (1948). A simple objective technique for measuring flexibility in thinking. The Journal of General Psychology, 39(1), 15–22.

Binz, M., & Schulz,

E. (2023). Using cognitive psychology to understand GPT-3.

Proceedings of the National Academy of Sciences, 120(6), e2218523120.

Bowden, E., & Beeman,

M. (2003). Normative data for 144 compound remote associate problems. Behavior Research Methods, Instruments, & Computers : A Journal of the Psychonomic Society, Inc, 35, 634–639. https://doi.org/10.3758/BF03195543 Butlin,

P. (2026). Higher-order representation in AI. Philosophy and the Mind Sciences, 7(1).

Camposampiero, G., Hersche, M., Wattenhofer, R., Sebastian, A., & Rahimi,

A. (2025). Can large reasoning models do analogical reasoning under perceptual uncertainty? arXiv preprint arXiv:2503.11207.

Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S., & Wu, C.-S. (2024). Art or artifice? large language models and the false promise of creativity. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 1–34.

Chan, C., Yim, Y., Zeng, H., Zou, Z., Cheng, X., Sun, Z., ···others. (2025). Xtom: Exploring the multilingual theory of mind for large language models. arXiv Preprint arXiv:2506.02461.

Channon, S., & Crawford,

S. (2000). The effects of anterior lesions on performance on a story comprehension test: Left anterior impairment on a theory of mind-type task.

Neuropsychologia, 38(7), 1006–1017. Chen, R., Jiang, W., Qin, C., & Tan,

C. (2025). Theory of mind in large language models:

Assessment and enhancement. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 31539–31558.

Chen, Y., Kirshner,

S. N., Ovchinnikov, A., Andiappan, M., & Jenkin,

T. (2025). A manager and an AI walk into a bar: does ChatGPT make biased decisions like we do?

Manufacturing & Service Operations Management, 27(2), 354–368.

Choi, J., Hong, Y., & Kim,

B. (2025). People will agree what I think: Investigating LLM' s False Consensus Effect. Findings of the Association for Computational Linguistics:

NAACL 2025, 95–126. Combs, K., Bihl, T., Howlett, S., & Adams,

Y. (2025). Zero-shot comparison of large language models (LLMs) reasoning abilities on long-text analogies.

De Langis, K., Park,

J. I., Le,

K. C., Schramm, A., Elfenbein, A., Mensink,

M. C., & Kang,

D. (2026). Strong memory, weak control: An empirical study of executive functioning in llms. Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 5971–5986.

DeCaro,

M.

S. (2016). Inducing mental set constrains procedural flexibility and conceptual understanding in mathematics. Memory & Cognition, 44(7), 1138–1148.

Delgado-Solorzano, C., DelaFlor, M., & Toxtli,

C. (2024). Assessing the Syllogistic Logic and Fact-Checking Capabilities of Large Language Models.

2024 IEEE International

Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics, 479–488. IEEE.

Desdevises,

J. (2025). The paradox of creativity in generative AI: high performance, humanlike bias, and limited differential evaluation. Frontiers in Psychology, 16, 1628486.

Diamond,

A. (2013). Executive Functions. Annual Review of Psychology, 64(Volume 64, 2013), 135-168. https://doi.org/https://doi.org/10.1146/annurev-psych-113011-143750

Dinu, A., & Florescu,

A.

M. (2024). An integrated benchmark for verbal creativity testing of llms and humans. Procedia Computer Science, 246, 2902-2911.

Artificial Theory of Mind in Large Language Models: Evidence, Definition, and Challenges

Theory of Mind (ToM) refers to the cognitive ability to understand and infer the mental states of oneself and others, such as beliefs, desires, and intentions. This capacity is fundamental to human social interaction and communication. With the rapid advancement of Large Language Models (LLMs), researchers have begun to explore whether these artificial systems possess a similar capability, often termed "Artificial Theory of Mind" (AToM).

1. Evidence for Theory of Mind in LLMs

Recent empirical studies have provided preliminary evidence suggesting that LLMs may exhibit behaviors consistent with Theory of Mind. Researchers have primarily utilized classic psychological paradigms, such as the False Belief Task, to evaluate these models.

[Figure 1: see original paper]

Initial testing indicated that advanced models, particularly those in the GPT-4 family, could successfully pass complex ToM tasks that were previously thought to be unique to humans. For instance, when presented with scenarios where a character holds a mistaken belief about the location of an object, these models can accurately predict the character' s behavior based on that false belief rather than the actual state of the world. Furthermore, LLMs have demonstrated the ability to recognize social faux pas, understand irony, and infer indirect speech acts, all of which require a degree of mental state attribution.

2. Defining Artificial Theory of Mind

Despite the impressive performance of LLMs on standardized tests, the definition of Artificial Theory of Mind remains a subject of intense debate. Unlike human ToM, which is rooted in biological evolution and embodied experience,

AToM in LLMs emerges from statistical regularities in vast amounts of text data.

We define Artificial Theory of Mind as the functional capacity of an AI system to represent and reason about mental states to predict or explain behavior. It is crucial to distinguish between "functional"ToM—the ability to solve ToM-related tasks—and "phenomenological" ToM, which would imply an internal subjective experience of understanding. Current evidence supports the existence of functional AToM in LLMs, while the question of internal representation remains an open research frontier.

3. Challenges and Limitations

The claim that LLMs possess Theory of Mind faces several significant challenges. These can be categorized into methodological, structural, and conceptual hurdles.

3.1 Methodological Challenges: The "Stochastic Parrot" Critique A primary concern is that LLMs may pass ToM tasks through pattern matching rather than

Eisape, T., Tessler, M., Dasgupta, I., Sha, F., Steenkiste, S., & Linzen,

T. (2024). A systematic comparison of syllogistic reasoning in humans and language models.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 8425–8444.

Fleming,

S.

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

Foote,

A. L., & Crystal,

J.

D. (2007). Metacognition in the rat. Current Biology, 17(6), 551–Gandhi, K., Fränken, J.-P., Gerstenberg, T., & Goodman,

N. (2023). Understanding social reasoning in language models with language models. Advances in Neural Information Processing Systems, 36, 13518–13529.

Gentner,

D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7(2), 155–170. https://doi.org/https://doi.org/10.1016/S0364-0213(83)80009-3 Gick,

M. L., & Holyoak,

K.

J. (1980). Analogical problem solving. Cognitive Psychology, 12(3), 306–355.

Gilhooly,

K. (2024). AI vs humans in the AUT: Simulations to LLMs. Journal of Creativity, 34(1), 100071.

Goto, D., Idei, H., Shiozuka, Y., & Ogata,

T. (2025). Performance of Large Language Models and Analysis of Responses in the Wisconsin Card Sorting Task.

2025 IEEE International

Conference on Development and Learning (ICDL), 1–7. IEEE.

Grant,

D. A., & Berg,

E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. Journal of Experimental Psychology, 38(4), 404.

Grassini, S., Grødem, S., Gunnarskog, T., & Sevre,

I.

T. (2025). Creativity in the Age of AI:

Comparing Human and Machine Performance Using Standardised Tests. In Human-

Computer Creativity: Generative AI in Education, Art, and Healthcare (pp. 49–65).

Springer. Gray, K., Anderson, S., Chen, E., Kelly, J., Christian, M., Patrick, J., ···Lewis,

K. (2019). "Forward Flow" : A New Measure to Quantify Free Thought and Predict Creativity.

American Psychologist, 74, 539–554. https://doi.org/10.1037/amp0000391 Guilford,

J.

P. (1967). Creativity: Yesterday, today, and tomorrow. The Journal of Creative Behavior, 1(1), 3–14. https://doi.org/10.1002/j.2162-6057.1967.tb00002.x Guzik,

E. E., Byrge, C., & Gilde,

C. (2023). The originality of machines: AI takes the Torrance Test. Journal of Creativity, 33(3), 100065.

Haase, J., Hanel,

P. H., & Pokutta,

S. (2025). Has the creativity of large-language models peaked?: An analysis of inter-and intra-llm variability. Journal of Creativity, 100113.

Hagendorff, T., Dasgupta, I., Binz, M., Chan,

S. C., Lampinen, A., Wang,

J. X., ···Schulz,

E. (2023). Machine psychology. arXiv Preprint arXiv:2303.13988.

Hagendorff, T., Fabi, S., & Kosinski,

M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nature Computational Science, 3(10), 833–838.

Hampton,

R.

R. (2009). Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms? Comparative Cognition & Behavior Reviews, 4, 17–

28. https://doi.org/10.3819/ccbr.2009.40002 Hao, G., Alexandre, F., & Yu,

S. (2025). Visual large language models exhibit human-level cognitive flexibility in the wisconsin card sorting test. IEEE Transactions on Cognitive and Developmental Systems.

Hao, L., You, W., & Xueling,

Y. (2025). Cognitive Biases in Artificial Intelligence:

Susceptibility of a Large Language Model to Framing Effect and Confirmation Bias.

Journal of Psychological Science, 48(4), 892–906. Happé,

F.

G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. Journal of Autism and Developmental Disorders, 24(2), 129–154.

Haq, S., Chhaya, N., Pandey, P., & Bhattacharya,

P. (2025). Is your LLM trapped in a Mental Set? Investigative study on how mental sets affect the reasoning capabilities of LLMs. arXiv Preprint arXiv:2501.11833.

Hatchuel, A., Masson,

P. L., & Weil,

B. (2017). C-K Theory: Modelling Creative Thinking and Its Impact on Research.

Haznitrama,

F. G., Ardi,

F. R., & Oh,

A. (2026). A Neuropsychologically Grounded Evaluation of LLM Cognitive Abilities. arXiv E-Prints, arXiv-2603.

Heit,

E. (2000). Properties of inductive reasoning. Psychonomic Bulletin & Review, 7(4), Hofstadter,

D.

R. (1995). Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought. (pp. ix, 518–ix, 518). Basic Books.

Holl-Etten,

A. K., Schnaderbeck, N., Kosareva, E., Prattke,

L. A., Krueger, R., Warner,

L. M., & Vetter,

N.

C. (2025). Applied Theory of Mind and Large Language Models-how good is ChatGPT at solving social vignettes? arXiv Preprint arXiv:2601.06032.

Holyoak,

K. J., & Thagard,

P. (1989). Analogical mapping by constraint satisfaction.

Cognitive Science, 13(3), 295–355. https://doi.org/10.1207/s15516709cog1303_1 Holzner, N., Maier, S., & Feuerriegel,

S. (2025). Generative AI and creativity: A systematic literature review and meta-analysis. arXiv Preprint arXiv:2505.17241.

Hou,

Z. J., Zhang,

B. A., Lu, Y., Baghel,

B. K., Brei, A., Lu, X., ···others. (2025).

CreativityPrism: A Holistic Benchmark for Large Language Model Creativity. arXiv Preprint arXiv:2510.20091.

Huang, L., Yu, W., Weitao, Zhong, W., Feng, Z., Wang, H., ···others. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), 1–55.

Huang, Y., Bie, B., Na, Z., Ruan, W., Lei, S., Yue, Y., & He,

X. (2025). An empirical study of the anchoring effect in llms: existence, mechanism, and potential mitigations. arXiv Preprint arXiv:2505.15392.

Huang, Z., Zhong, S., Zhou, P., Gao, S., Zitnik, M., & Lin,

L. (2025). A causality-aware paradigm for evaluating creativity of multimodal large language models. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Hubert,

K. F., Awa,

K. N., & Zabelina,

D.

L. (2024). The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks.

Scientific Reports, 14(1),

3440. Huff, M., & Ulakci,

E. (2025). Judgments of learning distinguish humans from large language models in predicting memory. Scientific Reports, 15(1), 35030.

Hwang, Y., Lee, D., Kang, T., Lee, M., & Jung,

K. (2026). When Wording Steers the Evaluation: Framing Bias in LLM judges. arXiv Preprint arXiv:2601.13537.

Jacowitz,

K. E., & Kahneman,

D. (1995). Measures of anchoring in estimation tasks.

Personality and Social Psychology Bulletin, 21(11), 1161–1166.

Ji-An, L., Mattar,

M. G., Xiong, H.-D., Benna,

M. K., & Wilson,

R.

C. (2025). Language models are capable of metacognitive monitoring and control of their internal activations.

ArXiv, arXiv-2505. Johnson-Laird,

P.

N. (1983). Mental models: Towards a cognitive science of language, inference, and consciousness. Harvard University Press.

Jones,

C. R., Trott, S., & Bergen,

B. (2024). Comparing humans and large language models on an experimental protocol inventory for theory of mind evaluation (EPITOME).

Transactions of the Association for Computational Linguistics, 12, 803–819.

Jung, C., Kim, D., Jin, J., Kim, J., Seonwoo, Y., Choi, Y., ···Kim,

H. (2024). Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 19794–19809.

Kahneman, D., & Tversky,

A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3(3), 430–454.

Kahneman, D., & Tversky,

A. (1984). Choices, values, and frames. American Psychologist, 39(4), 341.

Kepecs, A., Uchida, N., Zariwala,

H. A., & Mainen,

Z.

F. (2008). Neural correlates, computation and behavioural impact of decision confidence. Nature, 455(7210), 227–Knipper,

R. A., Knipper,

C. S., Zhang, K., Sims, V., Bowers, C., & Karmaker,

S. (2025). The Bias is in the Details: An Assessment of Cognitive Bias in LLMs. arXiv Preprint arXiv:2509.22856.

Koriat,

A. (1997). Monitoring one' s own knowledge during study: A cue-utilization approach to judgments of learning. Journal of Experimental Psychology: General, 126(4), 349–

370. https://doi.org/10.1037/0096-3445.126.4.349 Kornell, N., Son,

L. K., & Terrace,

H.

S. (2007). Transfer of Metacognitive Skills and Hint Seeking in Monkeys. Psychological Science, 18(1), 64–71. https://doi.org/10.1111/j.1467-9280.2007.01850.x Kumaran, D., Daw, N., Osindero, S., Velickovic, P., & Patraucean,

V. (2026). Causal Evidence that Language Models use Confidence to Drive Behavior. arXiv Preprint arXiv:2603.22161.

Lampinen,

A. K., Dasgupta, I., Chan,

S. C., Sheahan,

H. R., Creswell, A., Kumaran, D., ···Hill,

F. (2024). Language models, like humans, show content effects on reasoning tasks.

PNAS nexus, 3(7), pgae233.

Latif, E., Zhou, Y., Guo, S., Gao, Y., Shi, L., Nyaaba, M., ···Zhai,

X. (2025). Comparative evaluation of OpenAI O1 and human performance in higher order cognition. Scientific Reports.

Lee, D., Pruitt, J., Zhou, T., Du, J., & Odegaard,

B. (2025). Metacognitive sensitivity: The key to calibrating trust and optimal decision making with AI. PNAS Nexus, 4(5), pgaf133. https://doi.org/10.1093/pnasnexus/pgaf133 Lewis, M., & Mitchell,

M. (2024). Evaluating the robustness of analogical reasoning in large language models. arXiv Preprint arXiv:2411.14215.

Li, H., Zhang, G., Holme, P., Hu, S., & Wang,

Z. (2025). Large Language Models are NearOptimal Decision-Makers with a Non-Human Learning Behavior. arXiv Preprint arXiv:2506.16163.

Li, R., Zhu, C., Xu, B., Wang, X., & Mao,

Z. (2025). Automated creativity evaluation for large language models: A reference-based approach. arXiv Preprint arXiv:2504.15784.

Li, Y., Huang, Y., Wang, H., Cheng, Y., Zhang, X., Zou, J., & Sun,

L. (2024). Evaluating Large Language Models with Psychometrics. arXiv Preprint arXiv:2406.17675.

Li, Z., & Steyvers,

M. (2025). The Importance of Metacognitive Sensitivity in Human-AI Decision-Making. Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0). Retrieved from https://escholarship.org/uc/item/0fg7g94k Lior, G., Nachace, L., & Stanovsky,

G. (2025). WildFrame: Comparing framing in humans and LLMs on naturally occurring texts. arXiv Preprint arXiv:2502.17091.

Liu, H., Ding, Y., Fu, Z., Zhang, C., Liu, X., & Zhang,

Y. (2025). Evaluating the logical reasoning abilities of large reasoning models. arXiv Preprint arXiv:2505.11854.

Lou, J., & Sun,

Y. (2026). Anchoring bias in large language models: An experimental study.

Journal of Computational Social Science, 9(1), 11.

Lu, L.-C., Liu, M., Lu,

P. C., Tian, Y., Sun, S.-H., & Peng,

N. (2026). Rethinking creativity evaluation: A critical analysis of existing creativity evaluations. Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 6329–6352.

Luchins,

A.

S. (1942). Mechanization in problem solving: The effect of Einstellung.

Psychological Monographs, 54(6). Luchins,

A. S., & Luchins,

E.

H. (1959). Rigidity of behavior: A variational approach to the effect of Einstellung. (pp. xxv, 623–xxv, 623). Univer. Oregon Press.

Macmillan-Scott, O., & Musolesi,

M. (2024). (Ir) rationality and cognitive biases in large language models. Royal Society Open Science, 11(6).

Marchetti, A., Di Dio, C., Cangelosi, A., Manzi, F., & Massaro,

D. (2023). Developing ChatGPT' s Theory of Mind. Frontiers in Robotics and AI,

10. https://doi.org/10.3389/frobt.2023.1189525 Marchetti, A., Manzi, F., Riva, G., Gaggioli, A., & Massaro,

D. (2025). Artificial intelligence and the illusion of understanding: A systematic review of theory of mind and large language models. Cyberpsychology, Behavior, and Social Networking, 28(7), 505–514.

Marr,

D. (2010). Vision: A computational investigation into the human representation and processing of visual information. MIT press.

Masangkay,

Z. S., McCluskey,

K. A., McIntyre,

C. W., Sims-Knight, J., Vaughn,

B. E., & Flavell,

J.

H. (1974). The early development of inferences about the visual percepts of others. Child Development, 357–366.

Mayer,

R.

E. (1992). Thinking, problem solving, cognition. WH Freeman/Times Books/Henry Holt & Co.

Mednick,

S. (1962). The associative basis of the creative process. Psychological Review, 69(3), 220–232. https://doi.org/10.1037/h0048850 Milička, J., Marklová, A., VanSlambrouck, K., Pospíšilová, E., Šimsová, J., Harvan, S., & Drobil,

O. (2024). Large language models are able to downplay their cognitive abilities to fit the persona they simulate. Plos One, 19(3), e0298522.

Milligan, K., Astington,

J. W., & Dack,

L.

A. (2007). Language and theory of mind: Metaanalysis of the relation between language ability and false-belief understanding. Child Development, 78(2), 622–646.

Miyake, A., Friedman,

N. P., Emerson,

M. J., Witzki,

A. H., Howerter, A., & Wager,

T.

D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. Cognitive Psychology, 41(1), 49–100. https://doi.org/10.1006/cogp.1999.0734 Momennejad, I., Hasanbeig, H., Frujeri,

F. V., Sharma, H., Jojic, N., Palangi, H., ···Larson,

J. (2023). Evaluating cognitive maps and planning in large language models with cogeval. Advances in Neural Information Processing Systems, 36, 69736–69751.

Mondorf, P., & Plank,

B. (2024). Beyond accuracy: evaluating the reasoning behavior of large Language models–A survey. arXiv Preprint arXiv:2404.01869.

Musker, S., Duchnowski, A., Millière, R., & Pavlick,

E. (2025). LLMs as models for analogical reasoning. Journal of Memory and Language, 145, 104676.

Naeini,

S. A., Saqur, R., Saeidi, M., Giorgi, J., & Taati,

B. (2023). Large language models are fixated by red herrings: Exploring creative problem solving and einstellung effect using

the only connect wall dataset. Advances in Neural Information Processing Systems, 36, Newell, A., Simon,

H. A., & others. (1972). Human problem solving (Vol. 104). Prentice-hall Englewood Cliffs, NJ.

Nguyen,

H. M., & others. (2025). A survey of theory of mind in large language models:

Evaluations, representations, and safety risks. arXiv Preprint arXiv:2502.06470.

Nguyen,

J.

K. (2024). Human bias in AI models? Anchoring effects and mitigation strategies in large language models. Journal of Behavioral and Experimental Finance, 43, 100971.

O. Kelly, M., & Mandel,

D.

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

D.

E. (2025a). An anchoring effect in large language models. IEEE Intelligent Systems, 40(2), 23–26.

O' Leary,

D.

E. (2025b). Confirmation and specificity biases in large language models: An explorative study. IEEE Intelligent Systems, 40(1), 63–68.

Orrù, G., Piarulli, A., Conversano, C., & Gemignani,

A. (2023). Human-like problem-solving abilities in large language models using ChatGPT. Frontiers in artificial intelligence, 6, Osherson,

D. N., Smith,

E. E., Wilkie, O., Lopez, A., & Shafir,

E. (1990). Category-based induction. Psychological Review, 97(2), 185.

Ozeki, K., Ando, R., Morishita, T., Abe, H., Mineshima, K., & Okada,

M. (2024). Exploring reasoning biases in large language models through syllogism: Insights from the NeuBAROCO dataset. Findings of the Association for Computational Linguistics: ACL 2024, 16063–16077.

Pearl,

J. (2009). Causality: Models, Reasoning, and Inference. Cambridge University Press. https://doi.org/10.1017/cbo9780511803161 Perner, J., Leekam,

S. R., & Wimmer,

H. (1987). Three-year-olds' difficulty with false belief:

The case for a conceptual deficit. British Journal of Developmental Psychology, 5(2), Perner, J., & Wimmer,

H. (1985). “John thinks that Mary thinks that···” attribution of secondorder beliefs by 5-to 10-year-old children. Journal of Experimental Child Psychology, 39(3), 437–471.

Premack, D., & Woodruff,

G. (1978). Does the chimpanzee have a theory of mind?

Behavioral and Brain Sciences, 1(4), 515–526.

Qiu, Z., & Hu,

R. (2025). Deep Associations, High Creativity: A Simple yet Effective Metric for Evaluating Large Language Models. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 10870–10883.

Raven,

J.

C. (1941). STANDARDIZATION OF PROGRESSIVE MATRICES,

1938. British Journal of Medical Psychology, 19(1), 137–150. https://doi.org/10.1111/j.20448341.1941.tb00316.x
Ruan, K., Wang, X., Hong, J., Wang, P., Liu, Y., & Sun,

H. (2026). LiveIdeaBench:

Evaluating LLMs’ Divergent Thinking for Scientific Idea Generation with Minimal Context. Nature Communications.

Runco,

M. A., & Jaeger,

G.

J. (2012). The standard definition of creativity. Creativity Research Journal, 24(1), 92–96. https://doi.org/10.1080/10400419.2012.650092 Runco,

M. A., Turkman, B., Acar, S., & Alabbasi,

A.
B.

A. (2025). Examining the idea density and semantic distance of responses given by AI to tests of divergent thinking.

The Journal of Creative Behavior, 59(3), e1528. Sadhu, J., Khan,

A. A., Nawal, N., Basak, S., Bhattacharjee, A., & Shahriyar,

R. (2024).

Multi-tom: Evaluating multilingual theory of mind capabilities in large language models. arXiv Preprint arXiv:2411.15999.

Salmani-Zarchi,

M. M., Dousti,

M. J., Khaledi,

H. K., Mohajeri,

M. M., Vahabie, A.-H., & Faili,

H. (2025). From Neurotypical to Neurodiverse: Evaluating LLMs' Ability to Simulate Autistic Theory of Mind. Available at SSRN 5717345.

Sarıtaş, K., Tezören, K., & Durmazkeser,

Y. (2025). A systematic review on the evaluation of large language models in theory of mind tasks. arXiv Preprint arXiv:2502.08796.

Shaikh, A., Dandekar,

R. A., Panat, S., & Dandekar,

R. (2024). CBEval: A framework for evaluating and interpreting cognitive biases in LLMs. arXiv Preprint arXiv:2412.03605.

Shidara, K., Prem, P., Kim, J., Podlasek, A., Liu, F., Alaa, A., & Bernardo,

D. (2026).

Advances in LLM Reasoning Enable Flexibility in Clinical Problem-Solving. arXiv Preprint arXiv:2601.11866.

Simon,

H. A., & Kotovsky,

K. (1963). Human acquisition of concepts for sequential patterns.

Psychological Review, 70(6),

534. Song, Y., Qin, Y., Huang, R., Chen, Y., & Lin,

C. (2025). Legal text summarization via judicial syllogism with large language models. Journal of King Saud University Computer and Information Sciences, 37(5), 111.

St.

B.

T. Evans, J., Barston,

J. L., & Pollard,

P. (1983). On the conflict between logic and belief in syllogistic reasoning. Memory & Cognition, 11(3), 295–306. https://doi.org/10.3758/BF03196976 Sternberg,

R.

J. (1977). Component processes in analogical reasoning. Psychological Review, 84(4), 353.

Steyvers, M., & Peters,

M.

A. (2025). Metacognition and uncertainty communication in humans and large language models. Current Directions in Psychological Science, Stone, W., Stone, S., & Lindstedt,

J.

K. (2024). Beyond Scale: Deductive Reasoning Capabilities in Large Language Models Through the Lens of the Wason Selection Task. 2024 IEEE Western New York Image and Signal Processing Workshop (WNYISPW), 1-

4. IEEE.

Strachan,

J. W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., ···others. (2024). Testing theory of mind in large language models and humans. Nature Human Behaviour, 8(7), 1285-1295.

Street, W., Siy,

J. O., Keeling, G., Baranes, A., Barnett, B., McKibben, M., ···Dunbar,

R.

I. (2025). Llms achieve adult human performance on higher-order theory of mind tasks.

Frontiers in Human Neuroscience, 19,

1633272. Su, C., Li, H., Marques, M., Flint, G., Zhu, K., & Dev,

S. (2025). Limits of Emergent Reasoning of Large Language Models in Agentic Frameworks for Deterministic Games. arXiv. https://doi.org/10.48550/ARXIV.2510.15974 Sun, L., Yuan, Y., Yao, Y., Li, Y., Zhang, H., Xie, X., ···Stillwell,

D. (2025). Large Language Models show both individual and collective creativity comparable to humans.

Thinking Skills and Creativity, 57,

101870. Suri, G., Slater,

L. R., Ziaee, A., & Nguyen,

M. (2024). Do large language models show decision heuristics similar to humans? A case study using GPT-3.

5. Journal of Experimental Psychology: General, 153(4), 1066.

Suzgun, M., Gur, T., Bianchi, F., Ho,

D. E., Icard, T., Jurafsky, D., & Zou,

J. (2025).

Language models cannot reliably distinguish belief from knowledge and fact. Nature Machine Intelligence, 1–11.

Tenenbaum,

J. B., Griffiths,

T. L., & Kemp,

C. (2006). Theory-based Bayesian models of inductive learning and reasoning. Trends in Cognitive Sciences, 10(7), 309–318.

Tolman,

E.

C. (1948). Cognitive maps in rats and men. Psychological Review, 55(4), 189.

Tong, H., Yue, Z., Zhao, F., Lin, E., Jia, L., Chen, R., ···Zeng,

Y. (2026). CogToM: A Comprehensive Theory of Mind Benchmark inspired by Human Cognition for Large Language Models. arXiv Preprint arXiv:2601.15628.

Torrance,

E.

P. (1974). Torrance tests of creative thinking. Educational and Psychological Measurement.

Tourajmehr, A., Modarres,

M. R., & Yaghoobzadeh,

Y. (2025). Evaluating the Creativity of LLMs in Persian Literary Text Generation. arXiv Preprint arXiv:2509.18401.

Tversky, A., & Kahneman,

D. (1974). Judgment under Uncertainty: Heuristics and Biases:

Biases in judgments reveal some heuristics of thinking under uncertainty. Science, 185(4157), 1124–1131.

Ullman,

T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. arXiv Preprint arXiv:

2302.
2303. Ünlütabak, B., & Bal,

O. (2025). Theory of mind performance of large language models: A comparative analysis of Turkish and English. Computer Speech & Language, 89,

101698. https://doi.org/https://doi.org/10.1016/j.csl.2024.101698 Verma, M., Bhambri, S., & Kambhampati,

S. (2024). Theory of mind abilities of large language models in human-robot interaction: An illusion? Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, 36–45.

Vinchon, F., Gironnay, V., & Lubart,

T. (2024). GenAI creativity in narrative tasks:

Exploring new forms of creativity. Journal of Intelligence, 12(12), 125.

Wadhwa, M., Roy,

T. S., Lederman, H., Li,

J. J., & Durrett,

G. (2026). CREATE: Testing LLMs for Associative Creativity. arXiv Preprint arXiv:2603.09970.

Wallas,

G. (1926). The art of thought. Harcourt, Brace.

Wang, G., Wu, W., Ye, G., Cheng, Z., Chen, X., & Zheng,

H. (2025). Decoupling metacognition from cognition: a framework for quantifying metacognitive ability in LLMs. Proceedings of the AAAI Conference on Artificial Intelligence, 39, 25353–25361.

Wang, L., & Shen,

Y. (2024). Evaluating causal reasoning capabilities of large language models: A systematic analysis across three scenarios. Electronics, 13(23), 4584.

Wang, Q., Zhou, X., Sap, M., Forlizzi, J., & Shen,

H. (2025). Rethinking theory of mind benchmarks for llms: Towards a user-centered perspective. arXiv Preprint arXiv:2504.10839.

Wason,

P.

C. (1960). On the failure to eliminate hypotheses in a conceptual task. Quarterly Journal of Experimental Psychology, 12(3), 129–140.

Wason,

P.

C. (1968). Reasoning about a rule. Quarterly journal of experimental psychology, 20(3), 273–281.

Watson,

K. (2011).

D. Kahneman.(2011). Thinking, Fast and Slow. New York, NY: Farrar, Straus and Giroux. 499 pages. Canadian Journal of Program Evaluation, 26(2), 111–Webb,

T. W., Holyoak,

K. J., & Lu,

H. (2025). Evidence from counterfactual tasks supports emergent analogical reasoning in large language models. PNAS Nexus, 4(5), pgaf135.

Wimmer, H., & Perner,

J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children' s understanding of deception. Cognition, 13(1), 103–128.

Wu, W., & Deng,

W. (2025). Transitive Inference in Large Language Models and Prompting Intervention. ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5. IEEE.

Wu, Yaxiong, Liang, S., Zhang, C., Wang, Y., Zhang, Y., Guo, H., ···Liu,

Y. (2025). From Human Memory to AI Memory: A Survey on Memory Mechanisms in the Era of LLMs.

Retrieved from https://arxiv.org/abs/2504.15965 Wu, Yue, Tang, X., Mitchell,

T. M., & Li,

Y. (2023). SmartPlay: A Benchmark for LLMs as Intelligent Agents. arXiv. https://doi.org/10.48550/ARXIV.2310.01557 Xu, F., Lin, Q., Han, J., Zhao, T., Liu, J., & Cambria,

E. (2025). Are large language models really good logical reasoners? a comprehensive evaluation and beyond. IEEE Transactions on Knowledge and Data Engineering, 37(4), 1620–1634.

Xu, R., Sun, Y., Ren, M., Guo, S., Pan, R., Lin, H., ···Han,

X. (2024). AI for social science and social science of AI: A survey. Information Processing & Management, 61(3), Yin, X., Doost,

E. Z., Zhou, S., Yadav,

G. A., & Gorman,

J.

C. (2025). When Researchers Say Mental Model/Theory of Mind of AI, What Are They Really Talking About? arXiv Preprint arXiv:2510.02660.

Zhang, G., Ying, Y., Jiang, S., Liang, J., Yue, G., Fu, Y., ···Xiao,

Y. (2025). From Remembering to Metacognition: Do Existing Benchmarks Accurately Evaluate LLMs?

Findings of the Association for Computational Linguistics: EMNLP 2025, 13440–13457.

Zhang,

W. (2024). Empowering Multimodal Large Language Models for Solving Cognitive Puzzles. Proceedings of the 2024 2nd International Conference on Electronics, Computers and Communication Technology, 22–25.

Zhao, Y., Zhang, R., Li, W., & Li,

L. (2025). Assessing and understanding creativity in large language models. Machine Intelligence Research, 22(3), 417–436.

Zong, S., & Lin,

J. (2024). Categorical syllogisms revisited: A review of the logical reasoning abilities of LLMs for analyzing categorical syllogisms. Proceedings of the 1st Workshop on NLP for Science (NLP4Science), 230–239.

Note: Figure translations are in progress. See original paper for figures.