The Incomplete Bridge: How AI Research (Mis)Engages with Psychology

Han Jiang[◇*], **Pengda Wang**[♣*], **Xiaoyuan Yi**[♠†], **Xing Xie**[♠], **Ziang Xiao**[◇†]
[◇]Department of Computer Science, Johns Hopkins University
[♠]Microsoft Research Asia
[♣]Department of Psychological Sciences, Rice University
hjiang66@jh.edu pw32@rice.edu ziang.xiao@jhu.edu
{xiaoyuanyi,xing.xie}@microsoft.com

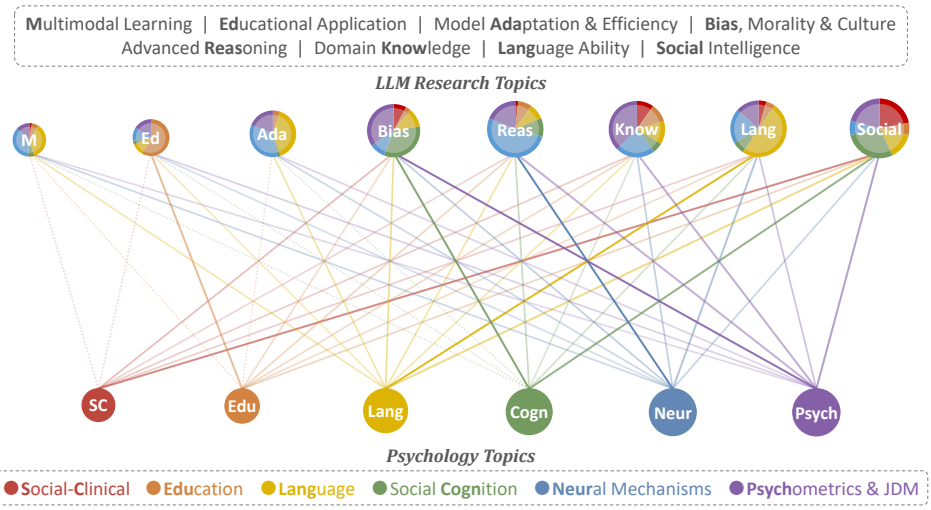

Figure 1: Bipartite Network of Citations from LLM Research Papers to Psychology Papers. *Note:* Pie charts show the distribution of psychology papers (six topics) cited by LLM research papers (eight topics). Circle size indicates the number of papers per topic; line opacity reflects citation frequency; dashed lines represent fewer than ten citations. Abbreviated topic labels are displayed for brevity; complete topic names are provided in §4.

Abstract

Social sciences have accumulated a rich body of theories and methodologies for investigating the human mind and behaviors, while offering valuable insights into the design and understanding of Artificial Intelligence (AI) systems. Focusing on psychology as a prominent case, this study explores the interdisciplinary synergy between AI and the field by analyzing 1,006 LLM-related papers published in premier AI venues between 2023 and 2025, along with the 2,544 psychology publications they cite. Through our analysis, we identify key patterns of interdisciplinary integration, locate the psychology domains most frequently referenced, and highlight areas that remain underexplored. We further examine how psychology theories/frameworks are operationalized and interpreted, identify common types of misapplication, and offer guidance for more effective incorporation. Our work provides a comprehensive map of interdisciplinary engagement between AI and psychology, thereby facilitating deeper collaboration and advancing AI systems.

[*] Equal contributions.
[†] Corresponding authors.

Contents

1 Introduction

In recent years, the rapid growth of artificial intelligence (AI) has enabled the development of more capable and innovative intelligent systems and has reshaped the way we study and conduct AI research. One of the most notable trends is toward research pluralism. Scholars increasingly recognize the importance of complementing traditional AI methodologies with broader approaches to further interpret, guide, and advance contemporary AI systems (e.g., Floridi, 2023; Lin and Dai, 2025). As a result, insights from the humanities and social sciences, particularly psychology, linguistics, cognitive science, and philosophy, are being integrated into AI research at an unprecedented scale and depth (e.g., Crawford, 2021; Lake et al., 2017; McCarthy et al., 2006).

One notable example of this interdisciplinary turn is the surge of interest in Large Language Models (LLMs; e.g., Gemini et al., 2024; Meta, 2025; OpenAI, 2024, 2025b). As AI systems progress from functioning primarily as interpreters to also becoming generative agents, LLMs have scaled up from smaller models that originally served as vehicles for language—the principal medium of human communication. With superior capabilities in natural language understanding and generation, these larger models have demonstrated remarkable performance across a wide range of downstream tasks and are increasingly integrated into practical applications across diverse sectors, including education (e.g.,Kasneci et al., 2023), healthcare (e.g.,Singhal et al., 2023), law (e.g.,Katz et al., 2024), scientific research (e.g.,Meyer et al., 2023), and commerce (e.g.,Li et al., 2024d). In doing so, LLMs have been profoundly redefining the modalities of knowledge acquisition and dissemination (e.g., Gao et al., 2024a). However, despite their impressive capabilities, the internal mechanisms of LLMs remain largely opaque (e.g.,Bommasani et al., 2022), rendering many of their exhibited behaviors unintended rather than well-understood or explainable (e.g.,Schaeffer et al., 2023; Wei et al., 2022b). LLMs are consequently characterized as *'black-box'* systems within both academic and industrial contexts, creating the challenging situation where users expect or steer what occurs without understanding why it occurs (e.g., Lipton, 2018; Rudin, 2019). Advancing LLM research urgently requires the development of systematic frameworks for evaluation, interpretability, and human-model interaction, which are essential for addressing foundational challenges related to reasoning mechanisms (e.g., Zhao et al., 2024a), capability boundaries (e.g., Chang et al., 2024), and alignment with human values and safety (e.g., Ouyang et al., 2022).

This is precisely where interdisciplinary research can play an important role in developing meaningful solutions. For example, the aforementioned research challenge concerning LLMs parallels a fundamental issue that psychology has grappled with since its inception. The discipline has progressed through the systematic observation and generalization of intelligent human behavior (e.g., Skinner, 1965; Watson, 1913; Wundt, 1904). In the absence of direct access to the underlying mechanisms of the human mind, psychologists have historically relied on rigorous experimental designs and theoretical modeling to describe, explain, predict, and influence cognitive and behavioral processes (e.g., Bunge, 2017; Gigerenzer, 1991; Shiffrin and Nobel, 1997). Drawing on this legacy, psychology's empirical tradition and sophisticated experimental paradigms offer a structured, systematic blueprint for advancing LLM research.

However, interdisciplinary collaboration is not without its challenges. When knowledge is transferred across disciplinary boundaries, researchers often encounter conceptual ambiguities and methodological tensions (e.g., Salter and Hearn, 1997; Huutoniemi, 2010). Terms may carry different meanings in different fields (e.g., differing interpretations of "attention" in LLM and psychology research), which can lead to misunderstandings, misinterpretations, and communication barriers in interdisciplinary research. Superficial understandings of complex theories may lead to misapplication (e.g., superficial use of motivation theory from psychology to AI design, which ignores human-specific factors like developmental context and lived experience), potentially overlooking critical context-specific factors and resulting in flawed designs or conclusions. In some cases, speculative or pseudo-scientific reasoning may inadvertently arise (e.g., equating the model's output diversity with human creativity), which can erode scientific credibility and mislead both academic and public audiences about the nature and capabilities of AI. Moreover, due to an insufficient or superficial understanding of theory, research may tends to repeatedly draw on a narrow set of well-known theories, favoring familiarity and accessibility over theoretical fit. This overreliance can crowd out alternative perspectives, stifle theoretical innovation, and reinforce conceptual blind spots in the field. All these challenges are deeply felt by AI researchers navigating interdisciplinary work, who often find themselves grappling with questions such as: Which areas of social science are relevant? What theories or frameworks from

these areas should we use? How should they be cited appropriately? And how can these concepts be meaningfully and responsibly integrated into technical research?

This paper explores how AI researchers are drawing on psychology literature in their work on LLMs, using this as a way to reflect on how current AI research engages with and integrates interdisciplinary theories and methods. Specifically, we address the following three core research questions and offer theoretical and methodological recommendations aimed at strengthening research pluralism in AI research, thereby guiding future exploration and facilitating better interdisciplinary research practices:

- **RQ1**: How is psychology research integrated into LLM research?
- **RQ2**: Which psychology theories/frameworks are most commonly used, and which remain underexplored in LLM research?
- **RQ3**: How is psychology research operationalized and interpreted in the context of LLM research?

We surveyed 25,843 LLM research articles and compiled a corpus of 1,006 papers that cited psychology research and were published in top-tier AI venues[3] between 2023 and 2025. From their references, we identified 2,544 psychology papers. By analyzing thematic patterns, we mapped the application of psychology research in LLM research, highlighting key areas of interdisciplinary overlap, and revealing potential research gaps.

Building on this foundation, our goal is to provide a rigorous, science-of-science analysis that maps the current intersection of AI and psychology, identifying emerging trends, critical gaps, and opportunities for impactful collaboration. By systematically assessing the landscape, we not only help researchers navigate this rapidly evolving space but also highlight areas where psychological insights can meaningfully inform AI development. This work is intended to foster responsible, well-informed interdisciplinary research, mitigate risks of conceptual misuse, and ultimately accelerate scientific progress in both fields. We believe that just as psychology has significantly advanced our understanding of human intelligence, it also holds the potential to play an important role in uncovering and guiding the behavioral mechanisms of AI systems. This study represents a key step toward that ambitious vision.

2 Preliminaries

2.1 Generative artificial intelligence and large language models

Generative AI is a subfield of AI that focuses on creating new content, such as text, images, audio, and video (Cao et al., 2025; Feuerriegel et al., 2024), representing a shift from merely interpreting input to generating novel outputs in response to user input. Among Generative AI systems, LLMs such as GPT (OpenAI, 2024, 2025b), Gemini (Gemini et al., 2024), and Llama (Meta, 2025) stand out as prominent text-based technologies. Similar progress is also observed in Multimodal Large Language Models (MLLMs; Yin et al., 2024), exemplified by LLaVA (Liu et al., 2024b), Claude 3 (Anthropic, 2024), and GPT-4V (OpenAI, 2025a), which further support visual content in both input and output. They have sparked a wave of research spanning their entire life cycle, from architecture and pre-/post-training to application and evaluation, which inherently engages a wide range of human-centered disciplines.

Existing LLMs are primarily decoder-only transformer models (Vaswani et al., 2017), in which the attention mechanism (e.g., Niu et al., 2021; Soydaner, 2022) is inspired by the concept of selective attention in cognitive science (e.g., Broadbent, 1958; Cherry, 1953). This architecture has greatly benefited from scaling (e.g., Alabdulmohsin et al., 2022; Kaplan et al., 2020), as first demonstrated by GPT-3 (Brown et al., 2020). The massive scale places significantly greater demands on computational resources and learning design, which typically occurs in two stages: pre-training, which involves learning from large-scale text corpora and aligns with the goals of corpus linguistics (Hunston, 2006); and post-training, which varies depending on specific objectives. Representative post-training methods include: 1) instruction tuning (e.g., Wei et al., 2022a; Zhang et al., 2024c), which enhances LLMs' generalization to unseen tasks; 2) alignment tuning (e.g., Wang et al., 2023d, 2024f), like Reinforcement Learning from Human Feedback (RLHF; Ouyang et al., 2022) and Direct Preference

[3]NeurIPS, ICLR, ICML, ACL, TACL, EMNLP, and NAACL

Optimization (DPO; Rafailov et al., 2023), which aligns LLMs with human intent and preferences; and 3) Parameter-Efficient Fine-Tuning (PEFT), such as prefix tuning (Li and Liang, 2021), prompt tuning (Lester et al., 2021), and Low-Rank Adaptation (LoRA; Hu et al., 2022), which enables effective adaptation of LLMs with a small subset of parameters being updated. PEFT is analogous to synaptic plasticity in neuroscience (Citri and Malenka, 2008), where only specific neural pathways are strengthened or weakened in response to new information.

Since their inception, and beyond language modeling and In-Context Learning (ICL; Dong et al., 2024), LLMs have demonstrated remarkable advancements in their core capabilities, including long-context modeling (e.g., Liu et al., 2024c; Yen et al., 2024), advanced reasoning (e.g., Huang and Chang, 2023; Liu et al., 2025a; Wei et al., 2023b), tool use (e.g., Qin et al., 2024), agency (e.g., Guo et al., 2024; Xi et al., 2023), and retrieval (e.g., Gao et al., 2024b; Wang et al., 2023b). These emergent capabilities are accelerating their widespread adoption, leading to the development of domain-specific LLMs in fields such as law (e.g., Lai et al., 2023b), economics (e.g., Horton, 2023), medicine (e.g., Singhal et al., 2023), education (e.g., Bernabei et al., 2023), and the arts (e.g., Wang et al., 2024a). Human-AI collaboration is simultaneously strengthening across the broad intersections between LLMs and diverse domains, resulting in many specialized AI assistants that are already in practical use (e.g., Microsoft, 2023). Despite notable progress, concerns about the social risks posed by LLMs continue to be raised (e.g., Bommasani et al., 2022; Street, 2024; Wang et al., 2023a; Weidinger et al., 2021), making evaluation and alignment crucial steps toward ensuring their responsible use and continuous improvement (Chang et al., 2024; Laskar et al., 2024). Numerous well-crafted benchmarks (e.g., Hendrycks et al., 2021; Srivastava et al., 2023; Zheng et al., 2023; Zhong et al., 2024) and dynamic approaches (e.g., Jiang et al., 2025; Zhu et al., 2024a) have been developed to evaluate LLMs from both holistic and targeted perspectives. Meanwhile, the objective of alignment is evolving from general human preference (e.g., Ji et al., 2023) toward personalized preferences (e.g., Salemi et al., 2024; Tan et al., 2024), steerability (e.g., Li et al., 2024b; Wang et al., 2024g), and social pluralism (e.g., Ashkinaze et al., 2025; Feng et al., 2024), reflecting a deepening exploration of LLMs' social impacts.

Nowadays, these models have long moved beyond being automatons confined to proof-of-concept experiments and are increasingly permeating human society, shaping both everyday life and various industries. As LLMs approach superhuman performance in certain professional tasks, there is growing interest in their subhuman or possibly non-human-like side. *Do* similar behavioral patterns or principles exist in LLMs (Gui and Toubia, 2023)? *What* are their internal mechanisms like (Han and Ji, 2025)? *How* can we learn from human cognition to overcome current limitations in efficiency (e.g., Alizadeh et al., 2024; Rostam et al., 2024), safety (e.g., Wolf et al., 2024; Yi et al., 2024), and social adaptiveness (e.g., Cuskley et al., 2024; Li and Qi, 2025)? The intensifying interconnection with human and emergent questions is pushing AI beyond computer science into the broader realm of social sciences.

2.2 Psychology

Psychology is an empirical science that systematically studies mental phenomena and behavioral mechanisms. It aims to uncover the underlying principles of how individuals perceive, think, feel, and behave in specific contexts (e.g., Colman, 2016; Dewey, 1892; James, 1892; Schacter et al., 2009). Since emerging from philosophy and physiology in the late 19th century, psychology has evolved from an early introspection-based approach to a methodology grounded in empirical research (e.g., Benjamin Jr, 2023; Goodwin, 2015; Schultz, 2013). Today, it encompasses a range of subfields, including cognitive, social, developmental, and clinical psychology (see American Psychological Association, n.d.).

As an interdisciplinary and methodologically diverse science, psychology investigates internal cognitive processes such as attention (e.g., Norman and Shallice, 1986), memory (e.g., Atkinson and Shiffrin, 1968; Baddeley, 2020), language (e.g., Chomsky, 2002; Pinker, 2003), and reasoning (e.g., Tversky and Kahneman, 1974), while also exploring how these processes are influenced by emotions (e.g., Damasio, 2006; LeDoux, 1998), motivation (e.g., Deci and Ryan, 2013, developmental stages (e.g., Piaget et al., 1952), and sociocultural environments (e.g., Henrich et al., 2010). Researchers employ various methods, such as experimental design (e.g., Reichardt, 2002), behavioral observation (e.g., Bakeman and Quera, 2011), surveys (e.g., Fowler Jr, 2013), neuroimaging techniques like functional magnetic resonance imaging (fMRI; e.g., Glover, 2011), and computational

modeling (e.g., Guest and Martin, 2021) to study psychological phenomena from multiple dimensions. These approaches emphasize the operational definition of variables and statistical inference in order to reveal systematic patterns underlying behavior and mental activity (e.g., Cohen, 1994; Kerlinger, 1966).

The core objectives of psychology can be delineated into four dimensions (e.g., Coon and Mitterer, 2013): description (the systematic observation and documentation of behavior and mental processes), explanation (the elucidation of underlying causes and mechanisms), prediction (the forecasting of future behavior based on theoretical frameworks), and intervention/influence (the ethically grounded facilitation of changes in psychological functioning and behavior). These objectives exhibit a noteworthy alignment with contemporary inquiries into LLMs. While LLMs are constructed through well-defined algorithmic architectures and trained on extensive datasets, many of their sophisticated capabilities (e.g., logical reasoning, code generation) have emerged not as explicit design features, but rather as emergent phenomena associated with increased model scale (e.g., Schaeffer et al., 2023; Wei et al., 2022b). Such phenomena underscore the current epistemic gap in our comprehension of LLMs' internal mechanisms: although we are accumulating observations of what these models can do, we still lack a systematic evaluation of their capabilities (e.g., Belinkov and Glass, 2019; Bommasani et al., 2022) and a clear understanding of the underlying reasons for their behaviors (e.g., Chang et al., 2024; Zhao et al., 2024a).

This epistemological asymmetry naturally invites interdisciplinary engagement. In particular, the theoretical paradigms and empirical methodologies developed within psychology may provide a productive lens through which to interrogate and interpret the behavior of LLMs (e.g., Kosinski, 2023; Lake et al., 2017). Psychology has historically played an important role in the development of AI, most notably during the early exploration of neural network theory, as exemplified by the perceptron model (Rosenblatt, 1958). More recently, psychological insights have continued to inform AI development; for example, attention mechanisms in advanced models (e.g., Vaswani et al., 2017) are conceptually inspired by research on human selective attention (e.g., Broadbent, 1958; Desimone et al., 1995; Treisman, 1964). Furthermore, recent studies have also demonstrated that insights from psychology research can significantly inform and enhance advancements in AI research (e.g., Liu et al., 2025d; Dong et al., 2025; Zhang et al., 2024b).

However, despite this growing body of interdisciplinary work, the broader potential of psychological science to contribute to contemporary AI remains substantial and, in many respects, underexplored. It is this largely untapped potential that motivates the present study. Our goal is to map the intersection between AI and psychology, identify trends, gaps, and opportunities for collaboration, and ultimately advance both fields. We hope to help researchers gain a clearer understanding of this domain and promote responsible interdisciplinary collaboration.

3 Analysis methodology

3.1 Data collection

We began our analysis with papers from seven prominent, peer-reviewed venues in machine learning, artificial intelligence, and natural language processing:

- *Annual Conference on Neural Information Processing Systems* (NeurIPS)
- *International Conference on Learning Representations* (ICLR)
- *International Conference on Machine Learning* (ICML)
- *Annual Meeting of the Association for Computational Linguistics* (ACL)
- *Conference on Empirical Methods in Natural Language Processing* (EMNLP)
- *North American Chapter of the Association for Computational Linguistics* (NAACL)
- *Transactions of the Association for Computational Linguistics* (TACL)

We collected papers presented at the 2023 and 2024 editions of these venues (except for NAACL, a biennial conference, from which we included only the 2024 edition) and added nine papers from Volume 13 of TACL in 2025 (N = 25,843). To ensure relevance to core LLM research areas, we only included papers that contained the terms *LLM* or *language model* in their title or abstract (N = 3,962).

Subsequently, we extracted citation data for the remaining papers using the Semantic Scholar Academic Graph (S2AG) API[4], which indexes over 214 million scholarly works across diverse scientific domains. From this citation dataset, we identified and retained only those referenced papers classified under the field of *Psychology* but not under *Computer Science*, according to Semantic Scholar's disciplinary tagging. LLM-related papers that did not cite at least one such psychology reference were excluded from the final sample (N = 1,006).

Following this multi-step curation process, our final dataset comprised 1,006 LLM research papers and 2,544 cited psychology reference papers.

3.2 Embedding and clustering

We employed the K-means clustering algorithm (Hartigan and Wong, 1979; Lloyd, 1982; MacQueen, 1967) to discern thematic groupings within corpora of LLM research papers and psychology reference papers, respectively. Specifically, we used the SPECTER model (Cohan et al., 2020) to generate embeddings for each paper. SPECTER is a transformer model trained on citation networks to produce document-level embeddings; it takes the title and abstract of a paper as input. Clustering was then performed across a range of cluster counts $K \in [4, 10]$, with the silhouette coefficient computed for each configuration to assess clustering quality. This procedure was repeated 50 times, and the value of K that yielded the highest average silhouette coefficient was selected as optimal.

This process yielded eight clusters for the LLM research papers and six for the psychology papers. The topic of each cluster was then inferred through a two-stage process: first, by summarizing the paper titles and abstracts within each cluster into five salient phrases using GPT-4o across ten runs, and second, by manually synthesizing these outputs into a concise, representative cluster label. The instruction template for summarization is provided in App. A, and the complete cluster names and descriptions can be found in §4.

3.3 Psychology theory/framework extraction and connection

To identify popular psychology theories and frameworks studied in the 2,544 psychology reference papers, we conducted the following three-step process: First, following the practice described in §3.2, we clustered the psychology papers within each of the six aforementioned clusters, resulting in 32 secondary clusters. Next, in each secondary cluster, we 1) randomly sampled 20 papers and summarized their titles and abstracts into five key phrases using GPT-4o, which was repeated ten times; and 2) hired domain experts to derive a cluster label, i.e., a research topic, and identify two to four primary psychology theories or frameworks based on the summaries from GPT-4o. Additionally, the experts were invited to suggest three more psychology theories or frameworks that are well-known in psychology but under-explored in LLM research for each primary cluster. Finally, we got a total of 96 popular psychology theories and frameworks, which were then linked to both psychology and LLM research papers.

To connect the identified theories and frameworks with the psychology papers in each primary cluster, we provided GPT-4.1 with the title and abstract of a paper, along with the list of popular theories/frameworks from a secondary cluster in each query. GPT-4.1 was then asked to determine whether the paper involves any of the listed theories or frameworks. The instruction template for the relevance judgment is provided in App. A. Once the relevant psychology papers were identified, we considered an LLM research paper to be associated with a given theory or framework if it cited any of the corresponding psychology papers. In this way, we calculated the citation count for each identified psychology theory or framework across the surveyed papers. For each primary cluster, the three most frequently cited psychology theories/frameworks and the three underexplored ones were selected for analysis in §5.2. The full list of secondary cluster names and the extracted popular psychology theories/frameworks are shown in Tables 5-10.

4 Clustering structure

As we mentioned in the previous section, we derive eight distinct LLM research clusters (shown in Fig. 2, left) and six distinct psychology clusters (shown in Fig. 2, right). In this section, we present

[4]https://www.semanticscholar.org/product/api

the names of the identified clusters and offer a brief overview of each. We begin with the LLM research clusters, followed by the psychology research clusters. For both sets, the corresponding topics are listed in ascending order based on the number of papers associated with each cluster.

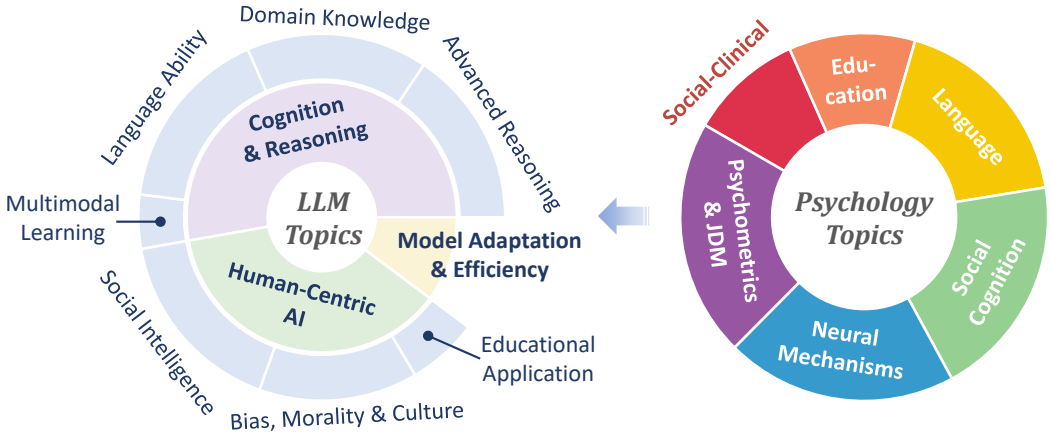

Figure 2: Illustration of Eight LLM Research Clusters and Six Psychology Clusters.
Note: The central angle of each segment indicates the proportion of research papers associated with its corresponding cluster. The LLM research clusters are organized in a two-tiered structure, with *Cognition & Reasoning*, *Human-Centric AI*, and *Model Adaptation & Efficiency* constituting the primary thematic layer. Abbreviated cluster labels are displayed for brevity; complete cluster names are provided in §4.

4.1 LLM research clusters

1) Multimodal Comprehension and Spatial Reasoning

- *Abbreviated as* ***Multimodal Learning***
- This cluster is characterized by the integration of modalities beyond text, such as images (e.g., Chakrabarty et al., 2023), audio (e.g., Kong et al., 2024), video (e.g., Liu et al., 2025c), and time series (e.g., Jin et al., 2024). Building on early research that primarily leveraged LLMs for the textual component of existing multimodal tasks, later directions including spatial reasoning (e.g., Wu et al., 2024b), concept binding (e.g., Li et al., 2024c), and multimodal generation (e.g., Zhen et al., 2024) have emerged with advances in LLMs and MLLMs. This branch of research has laid the foundation for embodied AI and more real-world applications.

2) Educational Applications and Pedagogical Alignment

- *Abbreviated as* ***Educational Application***
- This cluster explores how LLMs can be applied in educational settings, including educational material generation (e.g., Luo et al., 2024), assessment methods (e.g., Xiao et al., 2023), instructional design (e.g., Yin et al., 2023), and intelligent tutoring systems (e.g., Sonkar et al., 2023). The goal is to align models with sound pedagogical principles and ensure their effectiveness in supporting human teaching and learning.

3) Scalable and Efficient Algorithms for Learning and Inference

- *Abbreviated as* ***Model Adaptation & Efficiency***
- This cluster aims to improve the scalability and efficiency of LLM adaptation methods, encompassing pre-training (e.g., Dagan et al., 2024), post-training (e.g., Munos et al., 2024), and inference-time adaptation (e.g., Zhang et al., 2023c). The emphasis is on trade-offs between various aspects of the learning algorithms, such as overall performance versus computational cost (e.g., Dettmers et al., 2023) and alignment performance versus pre-training capabilities (e.g., Lin et al., 2024). In general, it focuses on relatively low-level algorithm designs and serves to accommodate a variety of expectations and use cases.

4) Bias Measurement, and Moral and Cultural Alignment and Evaluation

- *Abbreviated as **Bias, Morality & Culture***
- This cluster mainly addresses bias in LLMs (e.g., Manvi et al., 2024), which occurs as a consequence of complex interactions among morality (e.g., Abdulhai et al., 2024; Scherrer et al., 2023), culture (e.g., Li et al., 2024a; Shen et al., 2024b), ideology (e.g., Plaza-del Arco et al., 2024), and other factors. This line of research seeks to measure and mitigate harmful stereotypes by decomposing them into different social aspects and conducting analysis and alignment within each, so that LLMs can better respect diverse moral frameworks and cultural lenses during interaction.

5) Advanced Reasoning and Theory of Mind in Multi-Agent Systems

- *Abbreviated as **Advanced Reasoning***
- This cluster explores high-level reasoning abilities (e.g., Huang and Chang, 2023) that emerge with the upscaling of LLMs, including logical reasoning (e.g., Wang et al., 2024d), mathematical reasoning (e.g., Imani et al., 2023), and planning (e.g., Valmeekam et al., 2023). Another prominent subfield is theory of mind in multi-agent scenarios (e.g., Li et al., 2023; Wu et al., 2023), which enables LLMs to infer others' mental states—an ability essential for collaborative and socially intelligent systems. However, whether LLM reasoning constitutes merely structured, goal-directed pattern completion or resembles human-like thinking remains an open question.

6) Knowledge Utilization and Domain-Specific Applications

- *Abbreviated as **Domain Knowledge***
- This cluster enhances the ability of LLMs to manage and utilize knowledge, including resolving knowledge conflicts (e.g., Xu et al., 2024e), performing knowledge-grounded reasoning (e.g., Chen et al., 2024b), and conducting fact verification (e.g., Pan et al., 2023). Once the faithfulness of the information is ensured, the processed knowledge, both structured and unstructured, can be applied across domains such as medicine (e.g., Kim et al., 2024), law (e.g., Fei et al., 2024), and other areas where factual accuracy and specialized understanding are critical for practical applications.

7) Linguistic Competence, Multilingual Adaptation, and Text Generation Quality

- *Abbreviated as **Language Ability***
- This cluster focuses on the core capability of LLMs—language ability. Research primarily investigates basic linguistic processing (e.g., Kobayashi et al., 2024) and multilingual understanding (e.g., Tang et al., 2024; Zhang et al., 2023a,b), as well as more advanced language phenomena such as analogy (e.g., Wijesiriwardene et al., 2023), creativity (e.g., Gómez-Rodríguez and Williams, 2023), metaphor (e.g., Joseph et al., 2023; Wachowiak and Gromann, 2023), and ellipsis (e.g., Hardt, 2023; Testa et al., 2023). It aims to produce outputs that are grammatically correct, semantically coherent, and contextually appropriate.

8) Socially Aware and Emotionally Intelligent Dialogue Systems

- *Abbreviated as **Social Intelligence***
- This cluster centers on the social adaptiveness of LLMs—the ability to understand and navigate social situations effectively. An intelligent system should first avoid producing harmful content (e.g., Shaikh et al., 2023; Wei et al., 2023a), then develop an understanding of diverse social dynamics (e.g., Zhao et al., 2024b; Zhou et al., 2024b), enabling it to engage appropriately in social interactions (e.g., Kwon et al., 2024; Shao et al., 2023) and deliver emotionally resonant responses (e.g., Chen et al., 2023; Sabour et al., 2024), thereby fostering beneficial relationships between humans and AI in society.

4.2 Psychology research clusters

1) Social-Clinical Psychology of Mental Health and Intervention

- *Abbreviated as **Social-Clinical***
- This cluster explores the psychological foundations of mental health and clinical practice. It includes research on social influences (e.g., Liao et al., 2020; Meyer, 2003), therapeutic interventions (e.g., Fitzpatrick et al., 2017; Greimel and Kröner-Herwig, 2011), and the psychological

processes that underlie well-being (e.g., Diener et al., 1985, 2010), stress (e.g., Lazarus, 1966; Spitzer et al., 2006), and disorder (e.g., Cuijpers et al., 2010; Persson et al., 2019).

2) Learning, Teaching Design, and Educational Development

- *Abbreviated as **Education***
- This cluster focuses on how people learn and how educational environments can be optimized. It investigates instructional strategies (e.g., Kirschner et al., 2006; Miri et al., 2007), developmental pathways (e.g., Stipek and Iver, 1989; Zimmerman, 2000), and the cognitive mechanisms that support effective learning (e.g., Garner, 1987; Pintrich, 2002) and teaching (e.g., Kraft et al., 2018; Sullivan et al., 2014).

3) Language Comprehension, Pragmatic, and Psycholinguistic

- *Abbreviated as **Language***
- This cluster examines the psychological and cognitive processes involved in understanding and using language. Topics include real-time language comprehension (e.g., Ehrlich and Rayner, 1981; Levy, 2008), pragmatic inference (e.g., Goodman and Frank, 2016; Levinson, 2000), and the development (e.g., Berko, 1958; Oates and Grayson, 2004) and disorders (e.g., Boschi et al., 2017; Gorno-Tempini et al., 2011) of language.

4) Emotion, Morality, and Culture in Social Cognition

- *Abbreviated as **Social Cognition***
- This cluster investigates how emotions, moral reasoning, and cultural context shape our social understanding. It includes research on emotions (e.g., Moors et al., 2013; Scherer and Moors, 2019), empathy (e.g., Hoffman, 1996; Konrath et al., 2018), value systems (e.g., Schwartz, 2012; Graham et al., 2013), identity (e.g., Hegarty et al., 2018; Roccas and Brewer, 2002), and the ways people perceive and interact with others (e.g., Brown, 1986; Cuddy et al., 2009).

5) Neural and Cognitive Mechanisms of Learning and Creativity

- *Abbreviated as **Neural Mechanisms***
- This cluster focuses on the brain and cognitive systems that support learning, memory, and creative thinking. Research covers neuroimaging (e.g., Bookheimer, 2002; Kanwisher et al., 1997), computational modeling (e.g., Anderson, 2013; Tenenbaum et al., 2006), and the dynamic interplay between neural circuits and cognitive function (e.g., Baddeley, 2003; Wang et al., 2018).

6) Psychometrics, and Judgment and Decision-Making

- *Abbreviated as **Psychometrics & JDM***
- This cluster includes the study of psychometric measurement and the study of human decision processes. It includes scale development (e.g., Hamilton et al., 2016; John and Srivastava, 1999), modeling of cognitive biases (e.g., Nickerson, 1998; Tversky and Kahneman, 1981), and understanding how people assess risk (e.g., Lejuez et al., 2002; Mishra and Lalumière, 2011), probability (e.g., Bar-Hillel, 1980; Cosmides and Tooby, 1996), and outcomes (e.g., Hornsby and Love, 2020; Oliver et al., 1994).

For the sake of brevity, these clusters will hereinafter be referred to by their respective abbreviations.

5 Results

5.1 How is psychology research integrated into LLM research?

Once the citation analysis was finished, we were able to clearly observe the interrelationships between the eight LLM research clusters and the six psychology research clusters. The overall analysis results are presented in Fig. 1. Fig. 3 illustrates the temporal citation trends, while Fig. 4 provides a more detailed view of how each LLM research cluster has cited papers from the psychology clusters across different time periods. We observe the following patterns through our analysis.

Finding 1: LLM research has increasingly cited psychology research in recent years.

We first observed a growing trend in the incorporation of psychology research within the LLM literature over time. As shown in Fig. 3, the LLM research community has increasingly emphasized insights from psychology. This trend began around March 2023, when researchers started citing certain clusters of psychology research — notably the *Neural Mechanism*, *Language*, and *Psychometrics & JDM* clusters, which were among the earliest to receive attention. Subsequently, around July 2023, the volume of psychology-related citations in LLM research saw a marked increase. Later, by mid to late 2024, the overall growth in citation volume began to slow down.

We speculate that this emerging trend can be understood in several ways. First, the initial references to psychology in LLM research around March 2023 appear to coincide with the release of GPT-3.5-Turbo and GPT-4. This event may have sparked heightened academic interest in the inner workings of LLMs. At this early stage, researchers began drawing upon clusters most closely related to LLM mechanisms (*Neural Mechanism* and *Language*), as well as the cluster most relevant to model evaluation (*Psychometrics & JDM*). Around July 2023, the noticeable uptick in psychology-related citations may be linked to the release of open-source models like Llama2. The accessibility and flexibility of open-source models likely facilitated interdisciplinary collaboration, allowing researchers to more freely modify model architectures and behaviors to explore psychologically informed hypotheses and experiments. By the latter half of 2024, although citations of psychology in LLM research continued to rise, the rate of increase appeared to slow. This stabilization may suggest that core psychology research has largely been assimilated into the LLM research framework.

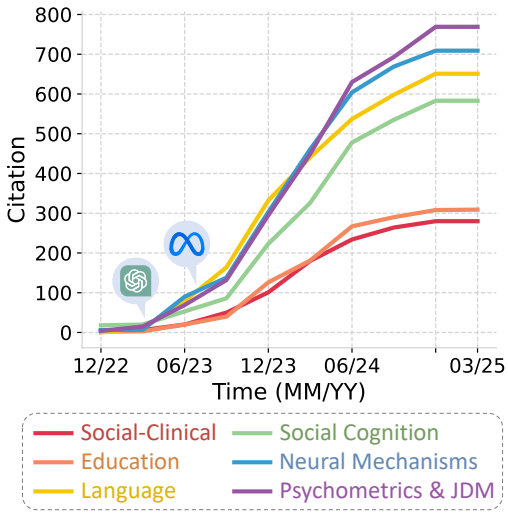

Figure 3: Overall Citation Flow from LLM Research Papers to Psychology Papers Over Time.

Finding 2: Psychology is broadly integrated into LLM research, with particular emphasis on certain clusters (i.e., *Psychometrics & JDM* and *Neural Mechanisms*).

We also found that psychology research is extensively referenced in LLM studies, with nearly all eight LLM research clusters drawing from multiple psychology clusters rather than a single domain (except for *Model Adaptation & Efficiency* $\leftarrow$ *Social-Clinical*; see Fig. 1). This suggests that LLM research engages with various areas of psychology, depending on the specific research questions or methodological needs.

Moreover, certain psychology research clusters are cited more frequently. As shown in Fig. 3, the distribution of citation frequency reveals a distinct hierarchical pattern: *Psychometrics & JDM* $\approx$ *Neural Mechanism* $>$ *Language* $>$ *Social Cognition* $\gg$ *Education* $\approx$ *Social-Clinical*. This likely reflects differences in methodological or conceptual alignment, with some research clusters being more closely aligned with the core objectives of current LLM research than others.

For example, the *Psychometrics & JDM* cluster contributes important tools for modeling and evaluating cognitive-like behaviors in LLMs. Foundations such as Classical Test Theory (CTT; Novick, 1966) and Item Response Theory (IRT; Lord, 1980) inform assessment frameworks, while work in judgment and decision-making (JDM) offers analogies for understanding LLM reasoning and uncertainty(e.g., Alabed et al., 2022; Placani, 2024). Similarly, the prominence of the *Neural Mechanism* cluster underscores the foundational role of neuroscience and cognitive psychology in shaping LLM research. Seminal contributions from these fields, such as Hebbian learning (Hebb, 1949) and early connectionist models like the perceptron (Rosenblatt, 1958), have directly influenced the design of neural architectures, including modern deep learning models like transformers (Alabdulmohsin et al., 2022; Kaplan et al., 2020).

In contrast, *Education* and *Social-Clinical* clusters are cited less frequently, which we speculate may be due to several reasons. First, research in these clusters often relies on long-term, large-scale human feedback collection, which can slow the pace of LLM-related advancements. Second, studies in these clusters often involve sensitive data, such as student or patient information, which raises

significant privacy concerns and makes data sharing and reuse more difficult due to strict ethical and legal constraints. Third, many contributions from these clusters are commonly submitted to the Human-Computer Interaction (HCI) community (Blandford, 2019; Fan et al., 2024a; Sanches et al., 2019), which is not included in this survey due to methodological differences.

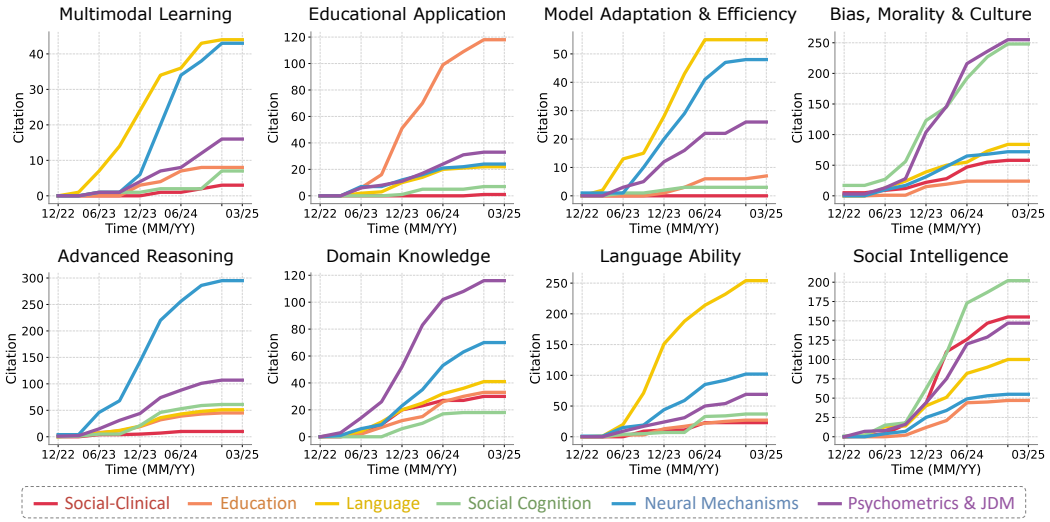

Figure 4: Citation Flow from LLM Research Papers to Psychology Papers Over Time.
Note: Each subfigure presents a grouping of research papers on LLMs organized by cluster, with colors indicating the corresponding clusters in the psychology literature. Abbreviated cluster labels are displayed for brevity; complete cluster names are provided in §2.

Finding 3: Different clusters of LLM research exhibit different tendencies in referencing psychology research.

After examining the overall patterns of how LLM research has cited psychology literature, we further explored the specific clusters within LLM research (see Fig. 4). Overall, different clusters of LLM research tend to favor different clusters of psychology research, reflecting variations in research focus.

For example, *Educational Application* shows a clear citation preference for *Education*, while *Advanced Reasoning* tends to favor citations from *Neural Mechanisms*. This pattern may be explained by the strong conceptual alignment between the LLM research cluster and the Psychology research cluster. Specifically, educational applications naturally draw on foundational work in educational psychology; whereas reasoning tasks tend to rely on insights from cognitive neuroscience to model complex inferential behavior, which can be traced back to neurons in Artificial Neural Networks (ANN; Jain et al., 1996).

Other clusters, such as *Model Adaptation & Efficiency* and *Social Intelligence* draw upon a substantially broader range of psychology clusters. This likely reflects the greater conceptual complexity inherent in constructs such as adaptation and awareness, which place higher demands on researchers to cite multiple aspects of psychology research. For example, *Social Intelligence* requires modeling human mental states and traits such as emotions (Ekman, 1992), cultural beliefs (Stivers et al., 2009), mental health (Elliott et al., 2018), and personality (John and Srivastava, 1999). This drives frequent citation of work from the *Social Cognition* and *Social-Clinical* psychology research clusters. At the same time, evaluating social awareness often involves extensive human-subject studies, which frequently result in citations of inter-rater reliability measures (Cohen, 1960; Fleiss, 1971; Spearman, 2010) from the *Psychometrics & JDM* cluster.

5.2 Which psychology theories/frameworks are most commonly used, and which remain underexplored in LLM research?

Building upon the analysis presented in §5.1, which examined the overall citation patterns of psychology research within the LLM literature, we now undertake a more granular investigation

Topic	Paper	Related Theory/Framework
Psychometrics & JDM	Measuring Nominal Scale Agreement Among Many Raters (Fleiss, 1971)	Classical Test Theory
Psychometrics & JDM	A Coefficient of Agreement for Nominal Scales (Cohen, 1960)	Classical Test Theory
Neural Mechanisms	Does the Chimpanzee Have A Theory of Mind? (Premack and Woodruff, 1978)	Theory of Mind
Psychometrics & JDM	The Proof and Measurement of Association between Two Things (Spearman, 2010)	Classical Test Theory
Neural Mechanisms	Catastrophic Forgetting in Connectionist Networks (French, 1999)	Complementary Learning Systems
Neural Mechanisms	Does the Autistic Child Have A "Theory of Mind"? (Baron-Cohen et al., 1985)	Theory of Mind
Psychometrics & JDM	A Technique for the Measurement of Attitudes (Likert, 1932)	Likert Scale
Psychometrics & JDM	Judgment under Uncertainty: Heuristics and Biases (Tversky and Kahneman, 1974)	Heuristics and Biases Program
Social Cognition	Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception (Wimmer and Perner, 1983)	Theory of Mind
Education	A Revision of Bloom's Taxonomy: An Overview (Krathwohl, 2002)	Bloom's Taxonomy

Table 1: Top 10 Most Cited Psychology Papers in LLM Research.

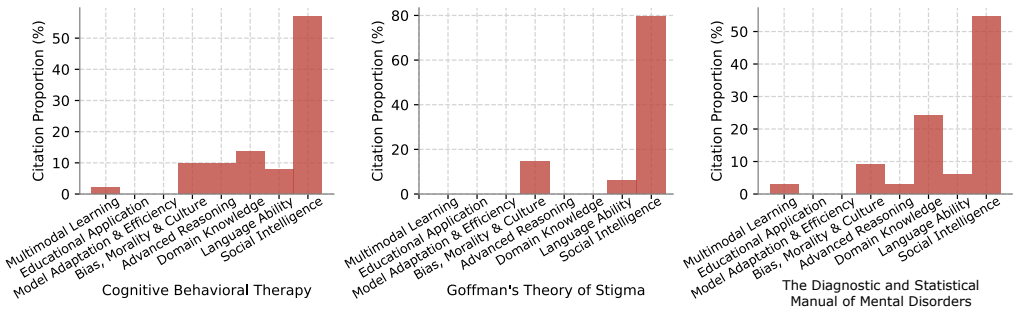

Figure 5: Citation Distribution for the Top Three Theories/Frameworks in the *Social-Clinical* Cluster Across Eight LLM Research Topics.

into how LLM research engages with specific psychology theories and frameworks within each identified psychology research cluster. For each cluster, we highlight both the most frequently cited theories/frameworks and those that remain underutilized or overlooked.

In addition to the cluster-specific analysis, we have identified the most influential psychology theories and frameworks across the entire LLM research landscape. By examining the top 10 most frequently cited psychology papers across all clusters (see Table 1), we find that the majority belong to the *Psychometrics & JDM* and *Neural Mechanisms* clusters, further supporting Finding 2 discussed in §5.1. The key theories represented in these influential works are Classical Test Theory (CTT) and Theory of Mind (ToM).

5.2.1 Social-clinical psychology of mental health and intervention

Popular theories/frameworks In the *Social-Clinical* cluster, the three most frequently referenced psychology theories/frameworks in the surveyed LLM research papers are *Cognitive Behavioral Therapy* (CBT; e.g., Beck, 2011), *Goffman's Theory of Stigma* (GTS; e.g., DeFleur, 1964), and *the*

Sub-Topic	Theory/Framework
Analysis and Application of Health Communication	Cognitive Behavioral Therapy ◀ The Belmont Report (Ethical Principals) Motivational Interviewing
Assessment Tools and Diagnostic Frameworks in Health	The Diagnostic and Statistical Manual of Mental Disorders ◀ Five Factor Model The Dark Triad
Therapeutic Processes, Intervention Methods, and the Therapeutic Relationship	Cognitive Behavioral Therapy ◀ The Working Alliance Motivational Interviewing
Stigma, Discrimination, and Health Disparities	Goffman's Theory of Stigma ◀ Minority Stress Model Intersectionality

Table 2: Subtopics and Corresponding Top Theories or Frameworks in the *Social-Clinical* Cluster. *Note:* Cell opacity represents citation frequency; black triangles indicate the three most frequently cited theories/frameworks.

Diagnostic and Statistical Manual of Mental Disorders (DSM; e.g., Association, 2013). Their citation distributions across the eight LLM research topics are shown in Fig. 5.

Cognitive Behavioral Therapy (CBT) is a psychotherapeutic framework that focuses on the interconnectedness of thoughts, emotions, and behaviors, aiming to help individuals identify and modify negative or maladaptive patterns. It has been demonstrated to be effective for a range of problems, including alcohol and drug use problems, marital problems, and severe mental illness. In this survey, CBT is the most frequently referenced theory/framework in the *Social-Clinical* cluster. LLM researchers primarily draw on paradigms from the CBT framework to develop models related to psychotherapy, resulting in 51 citations from the surveyed LLM research papers. For example, Wang et al. (2024c) used the CBT framework and LLMs to simulate virtual patients with various cognitive distortions, which could serve as a training tool for therapists to help them learn how to effectively formulate real cognitive models. Similarly, Xiao et al. (2024) also adopted the CBT framework and proposed an LLM-based mental enhancement model (empathetic dialogue system) for cognitive framing therapy. LLM research has also explored integrating LLMs into various stages of the CBT process. For example, Lissak et al. (2024) examined how LLMs could offer emotional support to queer adolescents, and Gabriel et al. (2024) evaluated the feasibility and ethical considerations of applying LLMs in mental health support.

Goffman's Theory of Stigma (GTS) is a theory that explores how individuals with attributes deemed undesirable by society experience social disapproval, exclusion, and discrimination. It emphasizes the role of societal norms and interactions in labeling individuals as deviant, leading to a spoiled social identity and altered self-concept. GTS has been influential in understanding the social dynamics of mental illness, physical disability, addiction, and other marginalized statuses, highlighting how stigma can affect access to resources, treatment engagement, and psychological well-being. LLM researchers have primarily drawn on GTS to explore whether LLMs exhibit bias and discrimination, and whether they amplify existing stigmas, resulting in 34 citations from the surveyed LLM research papers. For example, An et al. (2024) draws on the GTS that conceptualizes names as identity cues that function as social labels. By using gendered and ethnically marked names, they examine whether LLMs implicitly activate stereotypical associations tied to specific social groups. Similarly, Morabito et al. (2024) adopts the GTS point of stigma not as a discrete or isolated event, but as a structural and dynamic process. On this basis, this paper designs a dataset consisting of progressively intensified offensive language to model the escalation of stigmatization.

Diagnostic and Statistical Manual of Mental Disorders (DSM) is a standardized classification framework developed by the American Psychiatric Association for diagnosing mental health conditions. It provides clinicians with a common language and specific diagnostic criteria based on observable symptoms and clinical features. The DSM is widely used in research, clinical practice, and insurance reporting, and plays a central role in shaping the understanding, treatment, and categorization of mental disorders across diverse populations and settings. LLM researchers have

primarily drawn on the DSM framework to guide the application of LLMs in the mental health domain, resulting in 33 citations from the surveyed LLM research papers. The DSM provides clinical guidance, standardized symptom definitions, diagnostic labels, and decision-making criteria, thereby enhancing the scientific rigor, accuracy, and interpretability of LLM-based approaches. For example, Rosenman et al. (2024) leverages the DSM framework to enable LLMs to interpret unstructured psychological interviews for more accurate automated mental health assessments. Similarly, Kang et al. (2024), building on the DSM framework, integrates contextual information about symptoms to design a novel approach for LLM-based psychiatric disorder detection, aiming to reduce potential errors in automated symptom recognition.

Under-explored theories/frameworks In addition to the three theories/frameworks widely adopted by most LLM researchers, we also list three others that are closely related to *Social-Clinical* cluster but have received relatively little attention in current LLM studies. These theories have been extensively applied and have had a significant impact in the field of psychology. They also hold the potential to offer new perspectives and valuable insights for LLM research, making them well worth further exploration and consideration.

Biopsychosocial Model (BM) is a holistic framework that conceptualizes health and illness as the result of an interaction between biological, psychological, and social factors. It recognizes that mental and physical health are influenced not only by genetic or physiological processes, but also by emotions, thoughts, behaviors, relationships, and environmental contexts. BM is widely used in clinical assessment and treatment planning, especially in fields such as psychiatry, chronic pain management, and behavioral medicine, promoting a more integrated and person-centered approach to care. In LLM research, the BM can serve as a guiding framework for modeling user behavior and tailoring responses in a human-centered manner. By considering users' emotional states, cognitive patterns, and social contexts, LLMs can generate responses that are more empathetic, contextually relevant, and effective in addressing users' complex needs. This approach enhances user trust, satisfaction, and long-term engagement by aligning model behavior with the multifaceted nature of human experience.

Critical Race Theory (CRT) is a framework that examines how race and racism are embedded within social structures, institutions, and policies. It challenges the notion of racial neutrality and emphasizes that systemic inequality is maintained through laws, cultural narratives, and power dynamics that privilege dominant groups. CRT has been applied across disciplines such as education, public health, and sociology to highlight the lived experiences of marginalized communities and to advocate for structural change, equity, and social justice. In LLM research, CRT can serve as a lens to critically assess and mitigate biases in model outputs, training data, and deployment contexts. By incorporating CRT principles, researchers can better identify how LLMs might perpetuate racial stereotypes or inequities, and develop strategies to promote fairness, inclusivity, and accountability. This includes refining datasets, adopting more equitable evaluation metrics, and designing interaction protocols that center the voices and perspectives of historically marginalized users.

Health Belief Model (HBM) is a framework that explains health-related behaviors by focusing on individuals' perceptions of risk and benefits. It posits that behavior change is influenced by key factors such as perceived susceptibility to a health issue, perceived severity of the condition, perceived benefits of taking action, and perceived barriers to action. HBM has been widely applied in public health to design interventions that promote preventive health behaviors, such as vaccination, screening, and lifestyle modification, by addressing motivational and cognitive determinants of decision-making. In LLM research, the HBM can also serve as a theoretical framework for understanding and guiding user behavior. Researchers can leverage key components of the HBM, such as perceived susceptibility, perceived severity, perceived benefits, and perceived barriers, to design more persuasive and personalized dialogue strategies, thereby enhancing interaction quality and the model's ability to influence user behavior. By identifying users' motivations and concerns when responding to model-generated suggestions, LLMs can dynamically adjust their outputs to improve the adoption rate and trustworthiness of their recommendations.

5.2.2 Learning, teaching design, and educational development

Popular theories/frameworks In the *Education* cluster, the three most frequently referenced psychology theories/frameworks in the surveyed LLM research papers are *Self-Regulated Learning* (SRL; e.g., Graham and Harris, 1993), *Instructional Scaffolding* (IS; e.g., Beed et al., 1991), and

Sub-Topic	Theory/Framework
Cognitive, Social, and Developmental Factors of Academic Achievement	Dweck's Mindset Theory The Matthew Effect in Education Five Factor Model
Reading Comprehension and Vocabulary Development	Instructional Scaffolding ◀ Text Structure Theory Keyword Mnemonic Method
Pedagogy, Cognitive Processes, and Communication Strategies	Self-Regulated Learning ◀ Bandura's Social Cognitive Theory Dweck's Mindset Theory
Cognitive Science of Learning and Instructional Design	Bloom's Taxonomy ◀ Cognitive Load Theory Gagné's Conditions of Learning Theory

Table 3: Subtopics and Corresponding Top Theories or Frameworks in the *Education* Cluster. *Note:* Cell opacity represents citation frequency; black triangles indicate the three most frequently cited theories/frameworks.

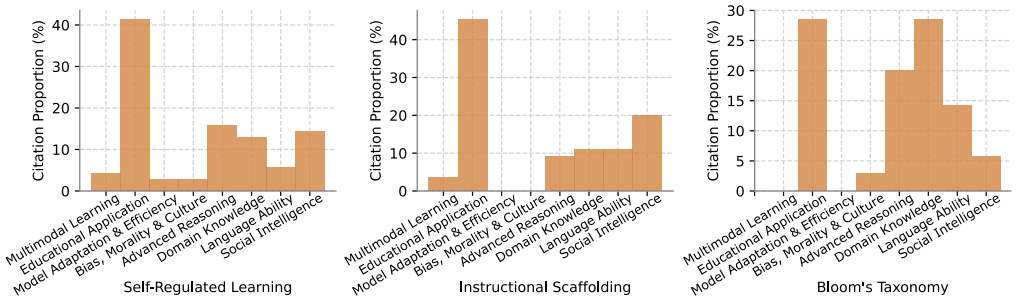

Figure 6: Citation distribution for the top three theories/frameworks in the *Education* cluster across eight LLM research topics.

Bloom's Taxonomy (BT; e.g., Bloom, 1956). Their citation distributions across the eight LLM research topics are shown in Fig. 6.

Self-Regulated Learning (SRL) is a theory that highlights the active role learners play in their own educational processes by setting goals, monitoring progress, and reflecting on outcomes. It emphasizes the cyclical interaction between cognitive, metacognitive, motivational, and behavioral components, enabling learners to strategically manage their learning environments and efforts. SRL has been shown to enhance academic performance, foster lifelong learning skills, and support students across diverse contexts, including online learning, special education, and higher education settings. In this survey, SRL is the most frequently referenced theory/framework in the *Education* cluster. LLM researchers have drawn on the SRL to inspire the design of self-improving LLM systems, resulting in 70 citations from the surveyed LLM research papers. The principles of SRL, especially self-monitoring and iterative feedback, have guided work in enabling LLMs to autonomously refine their outputs across domains, including problem solving (Gou et al., 2024; Wang et al., 2024e), fact-checking (Ying et al., 2024b; Yu et al., 2024a), code generation (Huang et al., 2024a), and data synthesis (Shi et al., 2024), thereby leveraging their reasoning capacity and internal knowledge structures. LLM also adopted the SRL theory to support education applications. For example, Macina et al. (2023) involved LLMs in the construction of dialog-based tutoring datasets to support reflective math learning, and Borges et al. (2024) explored how LLM-generated feedback can model and enhance SRL strategies in educational contexts.

Instructional Scaffolding (IS) is an educational framework that emphasizes the gradual support of learners as they develop new skills and understanding, with the ultimate goal of fostering independent competence. It involves the strategic use of guidance, prompts, and resources by instructors to bridge the gap between what learners can do alone and what they can achieve with assistance. IS has been shown to be effective in various learning contexts, including literacy development, problem-solving,

and complex conceptual learning across age groups and subject areas. In LLM research, IS has inspired the design of prompting strategies that position LLMs as instructional agents capable of guiding users through incremental reasoning. It received 55 citations from the surveyed LLM research papers. For example, Daheim et al. (2024), Sonkar et al. (2024), and Wang et al. (2024h) draw on IS principles to craft pedagogically informed prompts, enabling LLMs to lead users or other models through intermediate steps toward task completion. On the other hand, IS also offers a valuable conceptual lens for treating LLMs as adaptive learners that benefit from guided feedback. In this paradigm, LLMs mirror students who improve through iterative supervision, corrections, and example-based guidance. Such applications enhance the model's ability to internalize human preferences and refine performance over time. For example, Tian et al. (2024) and Wang et al. (2024e) implement feedback-driven training and refinement loops that emulate scaffolded learning processes, allowing LLMs to self-correct and align more closely with desired outcomes.

Bloom's Taxonomy (BT) is a hierarchical educational theory that categorizes cognitive skills into progressive levels, aiming to promote deeper learning and critical thinking. It serves as a foundational model for curriculum development, instructional design, and assessment strategies across educational settings, guiding learners from basic knowledge recall to advanced analytical and creative thinking. BT has been adopted in LLM research to guide the design and evaluation of model capabilities, and it was cited 35 times in the surveyed LLM research papers. By providing a structured hierarchy of cognitive complexity, BT enhances the interpretability, rigor, and educational alignment of LLM-based benchmarks and evaluation methodologies. For example, Alam et al. (2025), Cao et al. (2024), Wang et al. (2023c), and Yu et al. (2024a) reflect BT principles in benchmark construction, while Ying et al. (2024a) incorporates BT into dynamic benchmarking strategies. On the other hand, BT also serves as a conceptual lens for interpreting the internal structure of LLMs. For example, Liu et al. (2025b) explores how BT can shed light on the internal cognitive process of LLMs, and Wang et al. (2024b) draws on BT to investigate the underlying mechanisms of knowledge representation in language models.

Under-explored theories/frameworks We also list three other theories/frameworks that are closely related to the *Education* cluster but have received relatively little attention in current LLM studies.

Bronfenbrenner's Ecological Systems Theory (BEST) is a developmental theory that emphasizes the multiple layers of environmental influence on an individual's growth and behavior. It outlines how individuals are embedded within a series of interrelated systems, ranging from immediate settings like family and school (microsystem) to broader societal and cultural forces (macrosystem). The theory highlights the dynamic interactions between these systems and how changes in one layer can ripple through others, shaping developmental outcomes over time. BEST is widely applied in fields such as education, psychology, and public health to understand and support human development within context. In LLM research, BEST can serve as a framework for contextualizing user interactions by considering the multilayered influences on user behavior and preferences. By modeling users within their broader ecological environments, such as cultural norms, social relationships, and institutional contexts, LLMs can tailor their responses to better align with users' lived experiences, thereby enhancing relevance, empathy, and user engagement.

Simple View of Reading (SVR) is a theory that posits reading comprehension as the product of two primary components: decoding and linguistic comprehension. According to SVR, proficient reading occurs when individuals can accurately translate written symbols into spoken language (decoding) and effectively understand spoken language (comprehension). This model has been widely supported in empirical research and is used to inform assessments and interventions for reading difficulties, such as dyslexia and language impairment. In LLM research, SVR can serve as a framework for evaluating and enhancing models' reading comprehension abilities. By separately analyzing a model's decoding-like abilities (e.g., text recognition and parsing) and its linguistic comprehension abilities (e.g., understanding semantics and context), researchers can better identify specific strengths and limitations. This dual-component perspective can also guide the development of more targeted training strategies, improving both surface-level processing and deep understanding in LLM outputs.

Self-Determination Theory (SDT) is a theory that focuses on human motivation, emphasizing the role of innate psychological needs, autonomy, competence, and relatedness, in fostering self-motivated and healthy behavior. It has been widely applied across various domains, including education, healthcare, workplace settings, and psychotherapy, demonstrating effectiveness in enhancing well-being, intrinsic motivation, and sustained behavior change. In LLM research, SDT can serve as a

Sub-Topic	Theory/Framework
Narrative, Discourse and Meaning-Making	Schema Theory ◄ Conceptual Metaphor Theory Reader-Response Theory
Phonetics, Prosody, and Interaction in Spoken Communication	Embodied Cognition Theory Conversation Analysis Articulatory Phonology
Sociolinguistics, Culture, and Cross-Cultural Communication	Sapir-Whorf Hypothesis The Emergence Theory of Language Brown and Levinson's Politeness Theory
Pragmatic Inference and Information Processing in Dialogue	Grice's Theory of Implicature Rational Speech Act Brown and Levinson's Politeness Theory
Cognitive and Neural Foundations of Language Processing	Embodied Cognition Theory Construction-Integration Model The Simple View of Reading
Grammar, Lexicon, and Mental Representation	Connectionism vs. Symbolism ◄ Usage-Based Models of Language ◄ Generative and Universal Grammar
Computational Models of Language and Psychological Processes	Gricean/Post-Gricean Pragmatics Linguistic Inquiry and Word Count Conceptual Metaphor Theory

Table 4: Subtopics and Corresponding Top Theories or Frameworks in the *Language* Cluster. *Note:* Cell opacity represents citation frequency; black triangles indicate the three most frequently cited theories/frameworks.

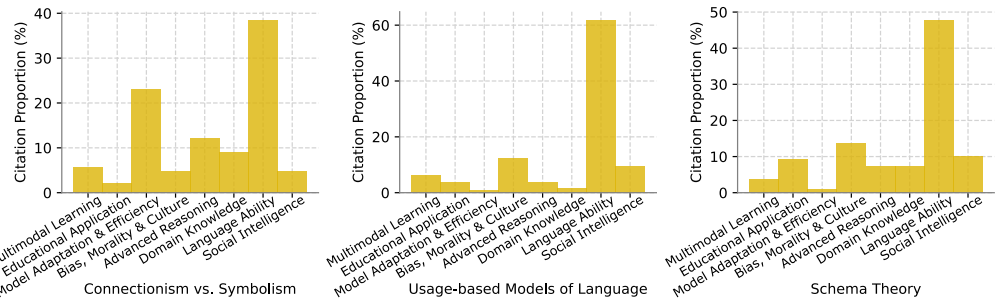

Figure 7: Citation distribution for the top three theories/frameworks in the *Language* cluster across eight LLM research topics.

guiding framework to promote user engagement and satisfaction by aligning model outputs with users' psychological needs. For example, LLMs can support autonomy by offering choices or allowing users to guide the direction of interactions, enhance competence by providing clear, constructive feedback, and foster relatedness through empathetic and personalized responses. By integrating SDT principles, LLMs can improve not only the effectiveness of user interactions but also long-term trust and continued use.

5.2.3 Language comprehension, pragmatic, and psycholinguistic

Popular theories/frameworks In the *Language* cluster, the three most frequently referenced psychology theories/frameworks in the surveyed LLM research papers are *Connectionism and Symbolism* (Faules and Alexander, 1978; Fodor and Pylyshyn, 1988), *Usage-based Models of Language* (von Mengden and Coussé, 2014), and *Schema Theory* (Graesser and Nakamura, 1982). Their citation distributions across the eight LLM research topics are shown in Fig. 7.

Connectionism and Symbolism represent two contrasting approaches in cognitive science aimed at explaining human thought and information processing. Connectionism relies on artificial neural networks to model mental processes, highlighting the role of distributed, parallel processing and learning from experience. Rather than explicit symbols, knowledge in connectionist models is stored in the patterns of connections and activation across the network. In contrast, Symbolism, also known as the rule-based approach, posits that cognition operates through the manipulation of discrete symbols according to formal rules, akin to programming languages or logical systems. It emphasizes structured representations and explicit reasoning mechanisms. While Connectionism offers strengths in pattern recognition, learning, and handling noisy or incomplete data, Symbolism excels at modeling rule-governed, high-level reasoning tasks. Contemporary cognitive science often seeks integrative frameworks that draw upon the complementary strengths of both paradigms.

Connectionism serves as a foundational theory for understanding and designing neural architectures in LLMs. This theory directly informs research on the internal mechanics and adaptive capabilities of LLMs. For example, work on model compression and pruning, such as Liu et al. (2024d) and Ma et al. (2023), leverages the connectionist notion of representational redundancy to identify and remove unnecessary parameters while preserving functionality. Similarly, generalization studies like Fan et al. (2024b), Yang et al. (2024), and Yu et al. (2024b) explore how internal patterns learned through training allow LLMs to extrapolate to novel linguistic tasks and inputs, reflecting classic connectionist learning dynamics. Moreover, work on emergent linguistic capabilities—such as de Varda and Marelli (2023) and Mitchell et al. (2023), illustrates how structured behaviors like syntax can arise from purely data-driven, neural processes, echoing one of Connectionism's core claims that complex cognition need not rely on explicit symbolic rules.

Symbolism, by contrast, has significantly influenced LLM research in areas that require explicit reasoning, compositional understanding, and interoperability. For example, Yuan et al. (2024) uses symbolic representations to enable structured musical instruction following, while Tennenholtz et al. (2024) investigates how LLMs encode modular and interpretable meaning representations. In formal linguistics, symbolic perspectives underpin works such as de Dios-Flores et al. (2023), Minixhofer et al. (2023), and Nair and Resnik (2023), which align LLM behavior with traditional linguistic theories involving morphology, syntax, and dependency structures. Furthermore, in mathematical and logical reasoning, rule-based symbolic models can help LLMs handle problems requiring multi-step, deterministic computation, as demonstrated by Imani et al. (2023) and Zheng et al. (2024).

Usage-based Models of Language (UBML) are theories that emphasize the role of actual language use in shaping linguistic knowledge and structure. Grounded in cognitive and functional approaches, UBML posits that language emerges from language users' repeated experiences with specific linguistic forms in meaningful contexts. These models highlight how frequency, context, and communicative function influence language acquisition, processing, and change. UBML has been applied to a wide range of theoretical and applied domains in linguistics, including grammar development, lexical patterning, language variation, and second language acquisition. LLM researchers have drawn on the UBML to understand and model how patterns of language usage influence LLM behavior and capabilities. For example, Zhang et al. (2023b) applies UBML theory to explain how LLM performance across different languages is shaped by the frequency of usage in training data. Similarly, Zeng et al. (2024) leverages the UBML perspective that language use shapes understanding to highlight LLMs' sensitivity to communicative strategies. In their work, LLMs are treated as human-like persuaders, and real-world rhetorical techniques, such as emotional appeals, appeals to authority, and logical reasoning, are employed to achieve jailbreaks. In addition, LLM researchers also use UBML theories as a basis for designing evaluation tasks aimed at assessing the LLMs' language abilities. For example, Wachowiak and Gromann (2023) draws on one of the core UBML theories, conceptual metaphor theory, to develop tasks that test whether LLMs can learn metaphorical mappings.

Schema Theory is a theory that posits that individuals organize knowledge into mental structures called schemas, which shape how they perceive, interpret, and respond to experiences. These schemas, developed through personal, social, and cultural narratives, guide meaning-making by filtering new information in accordance with existing beliefs and expectations. Schema theory explains how people construct coherence in stories, comprehend language, and derive significance from interactions by drawing upon pre-existing cognitive and cultural templates. Schema theory has been applied extensively in education and language learning to explain how learners use prior knowledge to comprehend new information. LLM researchers have drawn on schema theory to explore how LLMs emulate human-like reasoning and meaning construction through the activation of learned patterns.

For example, Sui et al. (2024) leverages the schema theory principle of generating coherent narratives based on pre-existing structures to reinterpret hallucinations in LLMs as a form of schema-like reasoning with potential narrative value. Similarly, Chen et al. (2024a) applies the schema activation mechanism to design a three-stage prompting framework (comprehend, associate, summarize) that simulates human cognitive processes to enhance generalization in multi-turn dialogue retrieval. In addition, schema theory has also been used by LLM researchers as an evaluative framework to assess model abilities. For example, Wicke and Wachowiak (2024) based on schema theory's embodiment-oriented perspective to examine whether LLMs and VLMs can demonstrate human-like intuitions about spatial schemas (e.g., support, containment, path) in the absence of sensorimotor grounding.

Under-explored theories/frameworks We also list three other theories/frameworks that are closely related to the *Language* cluster but have received relatively little attention in current LLM studies.

Predictive Coding (PC) is a theory that posits the brain as a predictive machine, constantly generating and updating models to anticipate sensory input. It emphasizes the interplay between top-down predictions and bottom-up sensory signals, where discrepancies drive learning and perception. PC has been influential in understanding perception, action, and cognition. In LLM research, PC may inform the development of adaptive and context-aware models by treating dialogue as a dynamic process of prediction and error correction. By modeling user inputs as sensory signals and the model's responses as top-down predictions, LLMs can iteratively refine their outputs based on user feedback and interaction history. This approach may enhance responsiveness, coherence, and personalization, enabling models to better align with user expectations and reduce communicative mismatches over time.

Hofstede's Cultural Dimensions Theory (HCDT) is a theory for understanding cultural differences through six key dimensions that influence how individuals perceive and interact with the world. It provides insights into national cultural values such as power distance, individualism vs. collectivism, masculinity vs. femininity, uncertainty avoidance, long-term orientation, and indulgence vs. restraint. HCDT serves as a valuable tool for navigating cross-cultural communication, enabling individuals and organizations to recognize and adapt to cultural nuances that affect communication styles, conflict resolution, leadership expectations, and collaboration in international or multicultural settings. In LLM research, HCDT can inform the development of culturally adaptive dialogue systems that are sensitive to users' cultural backgrounds. By incorporating insights from the six cultural dimensions, LLMs can tailor language, tone, and interaction strategies to align with users' communication preferences and expectations. This cultural alignment can enhance user engagement, reduce misunderstandings, and foster greater trust and effectiveness in diverse human-AI interactions. Moreover, HCDT can serve as an evaluative framework to assess whether LLM-generated content appropriately reflects or adapts to cultural norms, providing a systematic way to analyze model performance across different cultural contexts.

Grice's Cooperative Principle (GCP) is a framework in pragmatics, explaining how effective and meaningful communication is achieved in conversation. GCP posits that speakers typically aim to be cooperative by contributing appropriately to the communicative context. This principle is supported by four conversational maxims (Quantity, Quality, Relation, and Manner), which guide interlocutors to provide the right amount of information, to be truthful, relevant, and clear. GCP helps interpret implied meanings, identify conversational implicatures, and understand communication breakdowns when the maxims are flouted or violated. In LLM research, the GCP may serve as a foundational principle for designing more natural and contextually appropriate interactions. By aligning responses with the GCP framework, LLMs can enhance clarity, relevance, and trustworthiness in dialogue. Moreover, understanding when and how to strategically flout maxims (e.g., using understatement or irony) can help LLMs generate more nuanced and human-like communication. GCP may also serve as an evaluation framework for measuring the pragmatic appropriateness of LLM responses. This involves assessing how well the model adheres to conversational maxims and whether it produces implicatures in a contextually coherent manner.

5.2.4 Emotion, morality, and culture in social cognition

Popular theories/frameworks In the *Social Cognition* cluster, the three most frequently referenced psychology theories/frameworks in the surveyed LLM research papers are *Theory of Mind* (ToM; Premack and Woodruff, 1978; Towner, 2010), *Simulation Theory* (Shanton and Goldman, 2010), and

Sub-Topic	Theory/Framework
Collective Memory, Social Beliefs, and Self-Regulation	Bandura's Social Cognitive Theory Inoculation Theory Collective Memory Framework
Emotion Pragmatics, Culture, and Cross-Cultural Communication	Expectancy Violations Theory Empathy-Altruism Hypothesis Emotions as Social Information Model
The Foundations, Judgment, and Development of Morality	Moral Foundations Theory The "Big Three" of Morality Kohlberg's Stages of Moral Development
Persuasion, Deception, and Social Conflict	Theory of Mind ◄ Dual-Process Theory ◄ Inoculation Theory
Narrative, Empathy, and Psychological Influence	Simulation Theory ◄ Transportation Theory Experience-Taking
Personality Traits and Social Behavior	Five Factor Model The Dark Triad HEXACO Model of Personality
Social Identity, Stereotypes, and Cultural Values	Moral Foundations Theory Social Identity Theory Stereotype Content Model
The Theory, Perception, and Social Function of Emotion	Basic Emotion Theory Appraisal Theory of Emotion Circumplex Model of Affect

Table 5: Subtopics and Corresponding Top Theories or Frameworks in the *Social Cognition* Cluster. *Note:* Cell opacity represents citation frequency; black triangles indicate the three most frequently cited theories/frameworks.

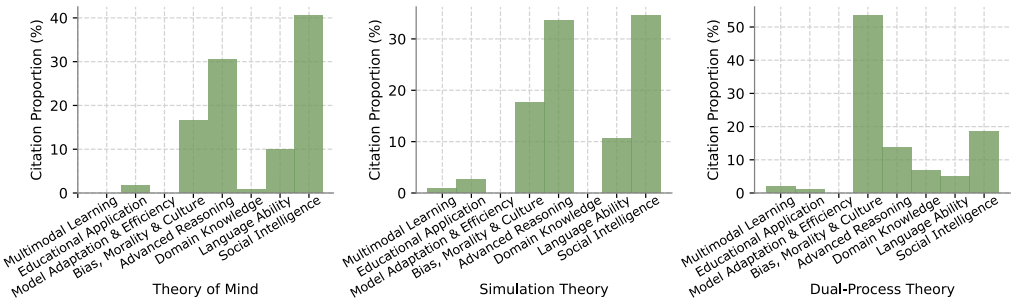

Figure 8: Citation distribution for the top three theories/frameworks in the *Social Cognition* cluster across eight LLM research topics.

Dual-Process Theory (Gawronski and Creighton, 2013). Their citation distributions across the eight LLM research topics are shown in Fig. 8.

Theory of Mind (ToM) is a psychology theory of the ability to understand other individuals by ascribing mental states to them. In ToM, others' beliefs, desires, intentions, emotions, and thoughts are recognized as potentially different from one's own. In other words, people use a theory of mind when analyzing, judging, and inferring others' behaviors; a well-functioning ToM is crucial for success in everyday social interactions. ToM is the most frequently referenced theory/framework in the *Social Cognition* cluster, and it was cited 121 times in the surveyed LLM research papers. In this cluster, LLM researchers have primarily drawn on ToM to interpret LLMs' social intelligence and their involvement in social interactions. For example, various benchmarks (e.g., Kim et al., 2023; Sabour et al., 2024) and tests (e.g., Shapira et al., 2023) were developed to examine LLMs' ToM capacity, which serves as a key indicator of their broader social intelligence. In addition, some

research has focused on LLMs' use of ToM in specific contexts. For example, Zhao et al. (2024b) evaluated LLM's ability to understand complex interpersonal relationships, and Xu et al. (2024a) incorporated personalized mental states into ToM assessment. There have also been analyses on ToM for goal-oriented dialogues. Bianchi et al. (2024); Chan et al. (2024); Lai et al. (2023a) explored how effectively LLMs can negotiate and persuade using ToM, and Wu et al. (2023) introduced a new deception mechanism within higher-order ToM reasoning.

Simulation Theory (ST) is a theory of how people understand others by engaging in a form of empathetic simulation. It posits that humans anticipate and make sense of others' behavior by internally simulating mental processes that, if enacted, would produce similar behaviors, such as intentional actions and emotional expressions. Compared to other theories of mind, ST draws more heavily on biological evidence. It has been applied across fields such as cognitive science, neuroscience, developmental psychology, and clinical research. In LLM research, ST forms a key foundation for AI empathy and other advanced social skills, and it received 113 citations from the surveyed LLM research papers. For instance, Qian et al. (2023) explored the role of ST in LLMs' empathetic responses, and Shen et al. (2024a) conducted an empirical analysis with LLMs examining the relationship between empathy and narrative style in storytelling. Their research has leveraged ST in empathy modeling, and Nie et al. (2023) examined the influence of ST on how LLMs make correct moral judgment. Furthermore, a sub-concept of ST called perspective-taking has been studied in LLMs. According to ST, it refers to the ability to understand a situation or concept from another person's point of view. Wilf et al. (2024) demonstrated that perspective-taking can effectively enhance LLMs' performance in mental state attribution, while Xu et al. (2024f) incorporated it into prompting and highlighted its significance in reducing bias and toxicity in LLMs. These findings indicate that ST may serve as a viable guide for developing more socially aware AI systems.

Dual-Process Theory (DPT) is a theory that explains how thought can arise in two distinct ways: System 1 and System 2. System 1 processes are implicit and unconscious, and may be influenced by persuasion or education; System 2 processes are explicit and conscious, typically requiring more time to adapt to different situations. These theories can be found in social, personality, cognitive, and clinical psychology, where the different modes of thinking are used to explain various phenomena. LLM researchers have primarily drawn on DPT to analyze and mitigate some social issues raised by these models, resulting in 103 citations among the surveyed LLM research papers. For instance, Sui et al. (2024) referred to DPT when accounting for hallucinations in LLMs, and Echterhoff et al. (2024); Koo et al. (2024) argued that both social and cognitive biases may stem from unconscious processes within LLMs, which parallel the System 1 processes described in DPT. Their work emphasizes the need for conscious audition to mitigate these problems. Meanwhile, inspired by DPT, researchers have found it effective to influence patterns of thinking and behavior in LLMs through the use of personas. Sun et al. (2024) assigned different visual personas to MLLMs and observed corresponding behavioral changes. Hu and Collier (2024) quantified the impact of assigned personas on perspective simulation. Liu et al. (2024a) further evaluated the resulting social bias and steerability induced by different persona assignments.

Under-explored theories/frameworks We also list three other theories/frameworks that are closely related to the *Social Cognition* cluster but have received relatively little attention in current LLM studies.

Cognitive Dissonance Theory (CDT) is a theory that focuses on the mental discomfort individuals experience when holding two or more conflicting cognitions, such as beliefs, attitudes, or behaviors. CDT posits that this dissonance motivates individuals to reduce the inconsistency, often by altering existing beliefs, justifying behaviors, or acquiring new information. CDT has been widely applied to understand processes like attitude change, decision-making, moral reasoning, and behavioral justification across various domains, such as consumer behavior, health psychology, and social dynamics. In LLM research, CDT can be used as a theoretical framework for understanding user resistance to model suggestions, particularly when those suggestions conflict with users' prior beliefs or intentions. By modeling and anticipating dissonant reactions, LLMs can be designed to offer responses that reduce psychological discomfort, for example, by providing justifications, alternative framings, or gradual nudges toward behavior change. Additionally, CDT can serve as an evaluation lens to assess how well LLM outputs align with users' cognitive states and to measure whether interactions successfully reduce dissonance over time, thereby enhancing long-term trust and acceptance.

Sub-Topic	Theory/Framework
Systems, Processes, and Brain Mechanisms of Memory	Episodic vs. Semantic Memory Complementary Learning Systems Baddeley's Model of Working Memory
Science of Learning in Minds and Machines	Complementary Learning Systems Bayesian Inference/The Bayesian Brain
Developmental Neuroscience of Mind and Brain	Executive Functions ◀ Theory of Mind ◀ Structure-Mapping Theory
Reasoning, Analogy, and Theory of Mind	Theory of Mind ◀ Theory-Theory ◀ Dual-Process Theory Structure-Mapping Theory
Cognitive Science Theories of Mental Architecture	Theory of Mind ◀ Dual-Process Theory Mental Model Theory of Reasoning

Table 6: Subtopics and Corresponding Top Theories or Frameworks in the *Neural Mechanisms* Cluster.
Note: Cell opacity represents citation frequency; black triangles indicate the three most frequently cited theories/frameworks.

Elaboration Likelihood Model (ELM) is a theory of persuasion that explains how individuals process and respond to persuasive messages through two distinct routes, the central route and the peripheral route. The central route involves careful and thoughtful consideration of the message content, typically occurring when the individual is motivated and able to engage in deep cognitive processing. In contrast, the peripheral route relies on superficial cues, such as the speaker's credibility, attractiveness, or emotional appeal, when motivation or ability to process is low. ELM provides a comprehensive framework for understanding attitude change, highlighting that the durability and strength of persuasion depend on the route through which it is achieved. It has been widely applied in areas such as marketing, health communication, and political campaigning. In LLM research, the ELM can serve as a framework for designing adaptive communication strategies based on user engagement levels. By assessing users' motivation and ability to process information, LLMs can tailor their responses to either follow the central route, providing detailed, logical arguments for highly engaged users; or the peripheral route, using concise, emotionally resonant cues for less engaged users. This approach may ensure that model outputs are better aligned with users' cognitive states. Additionally, the ELM can be used as an evaluative lens to assess the quality and impact of LLM-generated persuasive content, guiding improvements in user experience and behavioral outcomes.

Realistic Conflict Theory (RCT) is a theory that explains intergroup conflict as arising from competition over limited resources. It posits that when groups perceive that they are in direct competition for resources such as jobs, power, or territory, hostility and prejudice are likely to emerge. RCT emphasizes the role of tangible, real-world conflicts of interest in generating negative intergroup attitudes and behaviors. RCT has been applied to understand a variety of social phenomena, including ethnic tensions, discrimination, and political polarization. In LLM research, RCT may be used as a framework to analyze user interactions in competitive or zero-sum environments, such as online debates or resource allocation scenarios. By modeling how perceived intergroup threats influence communication patterns, developers can design LLMs that detect emerging conflicts and facilitate constructive dialogue. Additionally, RCT may inform the evaluation of LLM outputs in sensitive contexts by assessing whether responses exacerbate or mitigate perceived competition and group-based tensions.

5.2.5 Neural and cognitive mechanisms of learning and creativity

Popular theories/frameworks In the *Neural Mechanisms* cluster, the three most frequently referenced psychology theories/frameworks in the surveyed LLM research papers are *Executive Functions* (Diamond, 2013), *Theory of Mind* (ToM; Premack and Woodruff, 1978; Towner, 2010), and

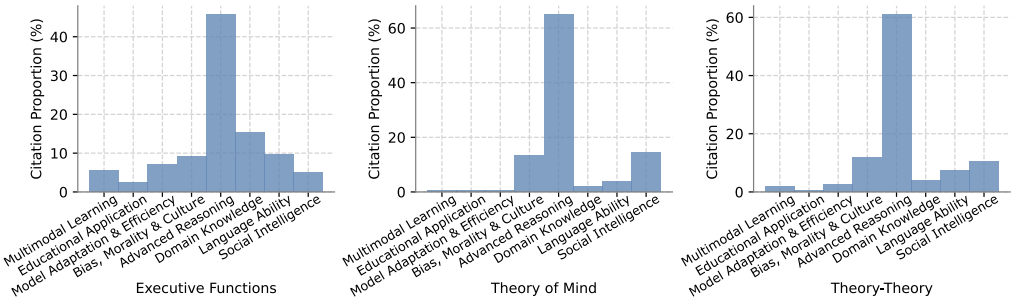

Figure 9: Citation distribution for the top three theories/frameworks in the *Neural Mechanisms* cluster across eight LLM research topics.

Theory-Theory (Ratcliffe, 2006). Their citation distributions across the eight LLM research topics are shown in Fig. 9.

Executive Functions (EFs) are a framework of cognitive processes that support goal-directed behavior, in which higher-order EFs require the coordinated use of multiple basic ones. All these functions develop gradually over the lifespan and can be improved at any point in a person's life, though they may be adversely affected by various affective factors. They play a fundamental role in people's actions and are deeply intertwined with domains such as mental health, social functioning, and academic achievement. EFs have been widely studied in education, clinical psychology, neuroscience, workplace settings, and public health and policy, particularly in terms of how they can be improved and maintained, as well as how they contribute to various forms of goal-directed behavior. EFs are the most frequently referenced theory/framework in the *Neural Mechanisms* cluster. LLM researchers have primarily drawn on EFs to enhance LLMs' capabilities for corresponding behaviors, resulting in 197 citations among surveyed LLM research papers. For example, higher-order EFs like problem-solving (e.g., Didolkar et al., 2024; Yao et al., 2023) and planning (e.g., Hao et al., 2023; Xie et al., 2024b) have become key areas in LLM reasoning. Meanwhile, some foundational EFs, though less extensively studied, have also been explored as a means to support the development of higher-level capabilities in LLMs. For instance, both working memory's essential role in LLMs' reasoning abilities (e.g., Wu et al. (2024b); Zhang et al. (2024a)) and cognitive flexibility's underpinning of their adaptive, context-sensitive behavior (e.g., Dong et al., 2023; Shao et al., 2023) have been empirically demonstrated. In addition to these main EFs, Ren and Xiong (2023) leveraged attention control to inhibit irrational shortcut learning, thus enhancing models' generalization.

Theory of Mind (ToM), the previously mentioned cognitive ability to attribute mental states, is supported by a dedicated network of brain regions and underlying EFs. Research on ToM in autism indicates that these specialized mechanisms can, in some cases, be selectively impaired while general cognitive function remains largely intact. Neuroimaging studies further evidence the view, showing that the medial prefrontal cortex (mPFC), the posterior superior temporal sulcus (pSTS), the precuneus, and the amygdala are involved in ToM-related activities (Shahaeian et al., 2011), such as social reasoning and decision-making. Therefore, in addition to its role in social reasoning, ToM has been adopted in LLM research to guide the exploration and enhancement of underlying reasoning mechanisms. In the psychology cluster of *Neural Mechanisms*, the 188 citations of ToM are heavily concentrated in *Advanced Reasoning* among the eight LLM research topics, exhibiting a distribution different from that shown in Fig. 8. A mainstream direction involves analyzing the mechanisms of LLM's social reasoning. For example, Huang et al. (2024b) measured the complexity of different ToM tasks for LLMs, drawing on cognitive load theory. Zhu et al. (2024b) linearly decoded LLMs' representations of their own and others' belief states from neural activations, adopting a connectionist perspective, while Jung et al. (2024) evaluated the precursory inferences for ToM in LLMs to further develop their ToM abilities, adopting a symbolic perspective. Another branch involves applying ToM to multi-agent interactions. This includes LLM collaboration, where the ToMs of different agents are integrated through synergy into a unified system (e.g.,Li et al., 2023; Wang et al., 2024h), and competition, where the ToMs of different agents are better differentiated and strengthened (e.g.,Du et al., 2024; Wu et al., 2024a; Xu et al., 2024c).

Theory-Theory (TT) is a theory concerning how humans develop an understanding of the outside world. While it shares with ToM the assumption that individuals possess a basic or naïve theory of

psychology (i.e., folk psychology) to infer others' mental states, TT is better understood as a broader framework for learning rather than a specific cognitive ability. It extends beyond reasoning about people and their viewpoints to include understanding mechanical devices and other non-agentive objects. TT has been widely studied in developmental psychology, education, cognitive modeling, social psychology, and domain-specific reasoning. From the perspective of TT, folk psychology, or the explanatory mental models in LLMs, is built through inductive reasoning. LLM researchers have primarily drawn on TT to enhance LLMs' capabilities in inductive reasoning and causal inference, resulting in 151 citations among the surveyed LLM research papers. For example, Wang et al. (2024d) proposed a pipeline for complex abstract hypothesis generation; Shani et al. (2023) and Suresh et al. (2023) mirrored TT to interrogate LLMs' latent structure of conceptual representations, thereby achieving concept awareness; and Jiayang et al. (2023) and Wijesiriwardene et al. (2023) evaluated LLMs across text analogies at various levels, ranging from words and sentences to metaphors and stories. These processes contribute to improved LLM performance in inductive reasoning tasks. Furthermore, Liu et al. (2023) and Nie et al. (2023) drew on TT when investigating LLMs' ability to derive cause-effect relationships.

Under-explored theories/frameworks We also list three other theories/frameworks that are closely related to the *Neural Mechanisms* cluster but have received relatively little attention in current LLM studies.

Levels of Processing Model (LPM) is a framework that posits memory retention is influenced by the depth at which information is processed. Rather than focusing on separate memory stores, the model emphasizes the continuum of processing levels, ranging from shallow (e.g., perceptual or structural features) to deep (e.g., semantic meaning and personal relevance). Deeper levels of processing lead to more durable and accessible memory traces. LPM has been influential in understanding encoding mechanisms and has implications for educational practices, memory enhancement strategies, and interventions for memory-related disorders. In LLM research, LPM holds great potential for enhancing the general learning process. Whether a similar mapping between levels of processing and memory duration exists in LLMs is worth investigating. If so, the three levels of processing (i.e., structural/visual, phonemic, and semantic) correspond to three modalities, which may inspire new learning algorithms for MLLMs. For example, LPM may serve as a guide for data administration at various training stages to mitigate forgetting or imbalance across different modalities. Moreover, if we further abstract the mapping in LPM, a conceptual parallel may emerge between the depth of human processing and the layers in deep neural models. By leveraging LPM-like mappings as heuristics in model adaptation, more PEFT methods become possible from an information processing perspective. Meanwhile, from a memory duration perspective, the personalization of LLMs can be effectively managed to balance steerability and stability.

Piaget's Stage Theory of Cognitive Development (PSTCD) is a theory that outlines how children's thinking evolves through a series of qualitatively distinct stages, each characterized by different cognitive abilities. PSTCD posits that children actively construct knowledge as they interact with their environment, progressing through four stages: sensorimotor, preoperational, concrete operational, and formal operational. Each stage represents a shift in how children understand and engage with the world, highlighting the importance of maturation and experience in cognitive growth. PSTCD has significantly influenced educational practices and our understanding of child development. In LLM research, PSTCD may serve as a valuable conceptual framework for modeling human-like learning trajectories. By incorporating the stage-based characteristics of cognitive development, researchers may design training curricula that progress from simple, concrete tasks to more abstract, logical reasoning, mirroring the natural evolution of human cognition. Furthermore, aligning interaction strategies with different cognitive stages allows LLMs to generate age-appropriate and educationally tailored responses, making them more effective for personalized learning environments. PSTCD also offers a structured lens for evaluating a model's cognitive maturity, guiding the design of benchmarks that reflect developmental reasoning skills.

Hebbian Theory (HT) is a theory that emphasizes the role of synaptic plasticity in learning and memory. HT posits that the repeated and persistent activation of one neuron by another strengthens the connection between them, thereby shaping neural networks. HT has been influential in understanding brain development, learning processes. In LLM research, HT may offer valuable inspiration for understanding and designing learning mechanisms. HT's core idea, neurons that fire together, wire together, has influenced the development of neural networks by introducing local learning rules and

Sub-Topic	Theory/Framework
Survey Design, Experimentation and Science Communication	Dual-Process Theory ◄
	Cognitive Aspects of Survey Methodology
	Cultural Consensus Theory
Measurement and Application of Psychometrics	Classical Test Theory ◄
	Item Response Theory
	Multitrait-Multimethod Matrix
Bias and Irrationality in Human Judgment	Heuristics and Biases Program ◄
	Rational Choice Theory/Game Theory
	Fuzzy-Trace Theory
Cognitive Models of Human Reasoning and Decision-Making	Heuristics and Biases Program ◄
	Causal Models/Causal Bayes Nets
	Evolutionary Psychology Approach to Reasoning

Table 7: Subtopics and Corresponding Top Theories or Frameworks in the *Psychometrics & JDM* Cluster.
Note: Cell opacity represents citation frequency; black triangles indicate the three most frequently cited theories/frameworks.

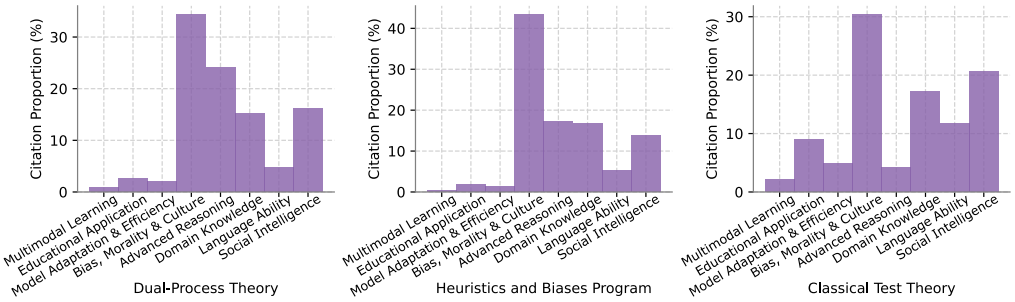

Figure 10: Citation distribution for the top three theories/frameworks in the *Psychometrics & JDM* cluster across eight LLM research topics.

concepts of synaptic plasticity. These ideas may support the exploration of more biologically plausible and interpretable models. Furthermore, HT emphasizes associative memory, may offer insights into how LLMs might enhance their capacity for long-term memory and contextual association. This may have implications for future architectures, such as neuromorphic computing and memory-augmented models, where energy-efficient and adaptive learning processes are increasingly important.

5.2.6 Psychometrics, and judgment and decision-making

Popular theories/frameworks In the *Psychometrics & JDM* cluster, the three most frequently referenced psychology theories/frameworks in the surveyed LLM research papers are *Dual-Process Theory* (Gawronski and Creighton, 2013), *Heuristics-and-Biases Program* (Tversky and Kahneman, 1974), and *Classical Test Theory* (CTT; Lord et al., 1968; Novick, 1966). Their citation distributions across the eight LLM research topics are shown in Fig. 10.

Dual-Process Theory (DPT), previously mentioned as the theory that partitions human cognition into two distinct types of processes, has significantly influenced studies of executive control, reward-based learning, and judgment and decision-making. According to some researchers, System 1 and System 2 do not operate as parallel systems (Walker, 2016). Typically, System 1 generates intuitive responses, which are then monitored and evaluated by System 2. However, System 2 does not always override System 1, especially under time pressure, cognitive load, or distraction. This has inspired LLM researchers to propose that not only social interactions but also a broader range of LLMs' behaviors can be moderated and steered through DPT and related frameworks. In the *Psychometrics & JDM* cluster, DPT is the most frequently cited psychology theory/framework, with 236 citations across the surveyed LLM research papers. For example, Yao et al. (2023) incorporated planning processes into general problem-solving, Dziri et al. (2023) used computation graphs for compositional tasks, and

Xu et al. (2024b) applied chain-of-thought reasoning in logical reasoning. Their research improves LLMs' decision-making by introducing a conscious, explicit guide to assist the unconscious, greedy processes. This comparison of LLMs' original cognitive processes to *System 1* in DPT is supported by evidence that LLMs can be easily persuaded (e.g., Xie et al., 2024a; Xu et al., 2024d). Goldstein et al. (2023) provides further evidence that LLMs' limited verification capability corresponds to *System 2*. Given this similarity, the trade-off between the two systems has invited discussions about consistency and uncertainty in LLMs (e.g., Jang and Lukasiewicz, 2023; Yona et al., 2024), particularly when dealing with knowledge conflicts.

Heuristics and Biases Program (HBP) is a research framework that investigates how people rely on heuristics to make decisions under uncertainty, and how these heuristics can lead to systematic errors or cognitive biases. Heuristics refers to the process by which humans use mental shortcuts to quickly arrive at judgments, decisions, and even solutions to complex problems. In the early 1970s, it became closely associated with cognitive biases through a series of experiments, demonstrating that people's intuitive judgments often deviated from normative rules. HBP informs research and practice in behavioral economics, clinical psychology, law, public policy, education, and organizational management. LLM researchers have mainly drawn on HBP to reveal how heuristics shape decision-making and how to mitigate related biases. HBP received 210 citations from the surveyed LLM research papers, most of which fall within the *Bias, Morality & Culture* cluster. For instance, Echterhoff et al. (2024) identified cognitive biases in LLMs under high-stakes scenarios and proposed a strategy for the models to mitigate their own human-like biases; Jiang et al. (2024) assessed whether LLMs possess genuine reasoning abilities or primarily depend on token bias. On the other hand, heuristics have been shown to be valuable for boosting the efficiency of LLMs in search (e.g., Gupta and Li, 2024; Yao et al., 2023) and reasoning (e.g., Bertolazzi et al., 2024; Pan et al., 2024). These heuristic approaches either mimic human heuristics or rely on computational strategies to balance performance and cost, while Zhou et al. (2024a) also validated that LLMs can automatically acquire task-specific heuristics from in-context demonstrations.

Classical Test Theory (CTT) is a psychometric theory concerned with predicting outcomes of psychological tests, such as item difficulty and examinee ability. It is based on the idea that a person's observed score on a test is the sum of a true score (the error-free score) and an error score. The aim of CTT is to understand and improve the reliability of psychological assessments, that is, to ensure test scores are precise, reproducible, and consistent across different testing conditions. LLM researchers have extensively adopted CTT in the design of evaluation methods, resulting in 145 citations among the surveyed papers. Many existing LLM evaluation methods *de facto* follow the logic of CTT implicitly, testing models with a range of items and reporting average scores (e.g., Manakul et al., 2023; Zheng et al., 2023). This is due to the straightforward assumptions of CTT, including the decomposition into true scores and errors, the linear and additive nature of the model, and the homogeneity of test items. Nonetheless, only a little research (e.g., Cao et al., 2024; Forde et al., 2024; Li et al., 2024e) formalizes error variance or true score modeling as CTT would. Moreover, some research has indicated that CTT would be less reliable for evaluating LLMs given its simple assumptions. For example, Xiao et al. (2023) observed noise in true score and error modeling, and proposed a testing framework to measure both the reliability and validity of NLG metrics. Kobayashi et al. (2024) proposed a benchmark that combines test items of varying evaluation granularity, aiming to mitigate inconsistencies across different tests. The context-sensitivity of LLMs and the nuanced nature of test items motivate the development of more context-aware, item-level, and dynamic measurements.

Under-explored theories/frameworks We also list three other theories/frameworks that are closely related to the *Psychometrics & JDM* cluster but have received relatively little attention in current LLM studies.

Lincoln and Guba's Evaluative Criteria (LGEC) is a framework for assessing the trustworthiness of qualitative research. LEGC emphasizes the importance of rigor through four key dimensions: credibility, transferability, dependability, and confirmability. These criteria serve as qualitative counterparts to the concepts of internal validity, external validity, reliability, and objectivity in quantitative research. The framework aims to ensure that qualitative findings are both believable and applicable, offering researchers a systematic approach to evaluating and enhancing the quality of their work. In LLM research, LGEC may guide the evaluation of model-generated qualitative outputs, such as LLM-as-judge for narrative responses, user reflections, or topic modeling. By applying

these criteria, researchers can assess whether the responses produced by LLMs are contextually relevant (transferable), logically consistent (dependable), and grounded in source data or reasoning (confirmable), thereby enhancing the trustworthiness and practical value of AI-driven qualitative analysis.

Prospect Theory (PT) is a theory that describes how individuals make decisions under conditions of risk and uncertainty, highlighting the psychological biases that diverge from rational choice. PT emphasizes that people evaluate potential losses and gains relative to a reference point, and that losses typically loom larger than equivalent gains (known as loss aversion). PT has been instrumental in explaining real-world decision-making patterns in areas such as finance, consumer behavior, and public policy. In LLM research, PT offers valuable insights for modeling and simulating human decision-making under uncertainty. By incorporating key principles such as loss aversion and reference dependence, LLMs can better reflect the psychological biases that influence human judgment. This is particularly useful in areas like dialogue generation, recommendation systems, and behavior prediction, where understanding users' risk preferences enhances personalization and realism. PT may also inform the design of reward functions in reinforcement learning from human feedback, enabling LLMs to align more closely with real-world human values and sensitivities. Furthermore, PT-guided framing strategies may improve the persuasive impact of generated content in domains like marketing, public policy, and healthcare communication.

Framing Theory (FT) is a theory that explores how information presentation influences individuals' perception and interpretation of events, issues, and messages. FT emphasizes the role of context, language, and emphasis in shaping meaning, guiding attention, and influencing emotional and behavioral responses. By highlighting certain aspects of a message while downplaying others, framing can significantly affect public opinion, decision-making, and social discourse. FT has been widely applied in media studies, political communication, public health campaigns, and social movement research. In LLM research, FT may offer valuable insights into how the presentation of prompts, context, and language influences model outputs. By highlighting the importance of emphasis, wording, and contextual cues, FT helps explain why different phrasings of the same question can lead to significantly varied responses from a model. This has implications for prompt engineering, bias detection, and user experience design. FT also aids in analyzing the latent frames within training data, which may introduce subtle biases into model behavior. Furthermore, FT may help researchers to explore public reactions under various narrative styles and assess emotional tone. Such as research on persuasive strategies in human-AI interaction can draw on this theory to examine how language framing influences user compliance and perception.

5.3 How is psychology research operationalized and interpreted in the context of LLM research?

In §5.2, we provided an overview and analysis of the psychology theories/frameworks cited in LLM research. Building on that foundation, this section further explores how LLM research concretely operationalizes and interprets the psychology literature, theories, and frameworks it references. Unlike the more macro-level overview in §5.2, the focus here is on the specific ways LLM research applies these psychological insights in practice, including potential misapplications or oversights.

Given the varying theoretical depth and scope of different psychology research, it is inevitable that LLM research exhibits diverse approaches to its application. Therefore, we adopt a case study approach to closely examine how the particular theory/framework is cited and used. Considering that ToM is one of the most commonly referenced psychology concepts in LLM research, we select ToM as a central case for in-depth analysis.

Although LLM research displays considerable variation in how it references different psychology theories, there are common patterns of misapplication in operationalization and theoretical understanding. Through the analysis of the ToM case, we aim to reveal these shared issues and offer insights for more rigorous and accurate incorporation of psychology research into LLM studies.

5.3.1 Case study: Theory of Mind

As introduced in §5.2, Theory of Mind (ToM) refers to the capacity of individuals to attribute mental states such as beliefs, intentions, knowledge, and emotions to others, recognizing that these states may differ from their own. This concept has become an important focus in LLM research, as it

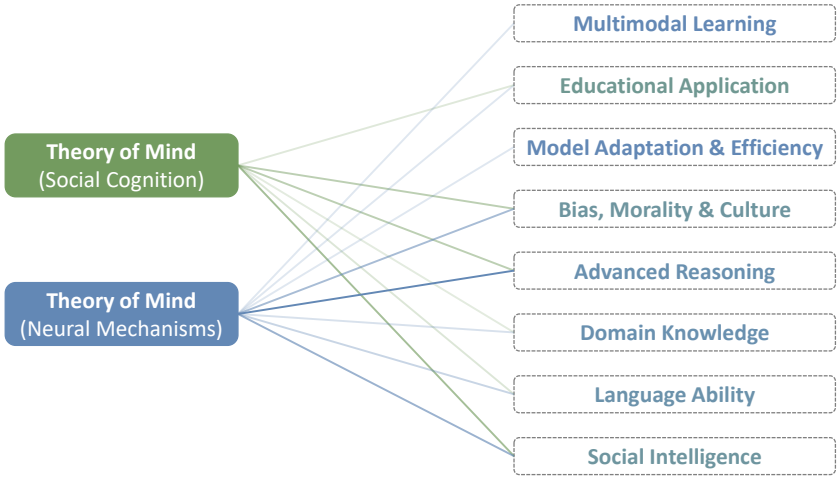

Figure 11: Bipartite Network of Citations Linking LLM Research Papers to Psychology Papers on Theory of Mind.
Note: Line opacity reflects citation frequency, and the color of LLM research topics indicates the proportion of citations to psychology papers in the two clusters.

reflects the models' potential to comprehend and reason about complex human mental states, which is considered essential for the development of more advanced and human-like artificial intelligence systems. As discussed in §5.2, ToM is among the three most influential theories/frameworks cited in LLM research, drawing from the *Social Cognition* and *Neural Mechanisms* clusters within psychology research. We conducted a comprehensive analysis of LLM papers that reference ToM-related works from these two clusters.

ToM-related psychology papers within the *Social Cognition* and *Neural Mechanisms* clusters show clear differences in research orientation and methodological approaches. Research in the *Social Cognition* cluster primarily focuses on the role of ToM in social interactions, emphasizing its relationships with abilities such as emotion understanding, social reasoning, empathy, and moral judgment. These studies often employ behavioral experiments, questionnaires, or situational tasks, highlighting the influence of developmental processes and social environmental factors on ToM. In contrast, research in the *Neural Mechanisms* cluster is more concerned with the biological foundation of ToM, exploring activation patterns in relevant brain regions (such as the prefrontal cortex and the temporoparietal junction) during ToM tasks. These studies commonly use techniques like functional magnetic resonance imaging (fMRI) and electroencephalography (EEG), following a neuroscience-oriented paradigm and focusing more on revealing the neural structure and functional mechanisms of ToM. Therefore, the way LLM research references and uses ToM-related papers from different clusters also tends to differ.

When LLM research references studies from the *Social Cognition* cluster, it primarily draws on established experimental tasks and research paradigms from that field as tools to evaluate the model's "ToM-like" abilities. For example, researchers often borrow tasks like the False Belief Task (e.g., the classic Sally-Anne test) to assess whether a model can distinguish between a character's perspective and reality. Other common tests like the Smarties test (first-order) and the Ice Cream Van test (second-order) are also used to evaluate whether the model can maintain mental state modeling across multiple dialogue turns. Additionally, some LLM studies adapt situational attribution tasks, narrative comprehension tasks, and social reasoning tasks to examine whether a model can grasp implicit intentions, emotional shifts, or social norms of characters. These tasks offer a structured and comparable framework for assessing LLMs.

When LLM research references studies from the *Neural Mechanisms* cluster, the primary focus is on how ToM is supported at the neural level, as well as the mechanisms and biological foundations underlying how some individuals process others' mental states. This includes examining the functional differentiation of brain regions such as the prefrontal cortex and temporoparietal junction in tasks like attribution, behavior prediction, and emotion understanding. Inspired by these studies, LLM researchers design new architectures for LLMs, develop multi-agent systems/frameworks, and seek

to explain model behaviors. For example, some LLM researchers introduce processes analogous to ToM by incorporating concept-level representations during training, which they want to allow models to develop an understanding of concepts prior to engaging in tasks like language generation or comprehension. Another example is LLM researchers inspired by the structure of social cognitive networks in neuroscience to structure the multi-agent systems, where each agent simulates the mental state of a specific role.

Common misapplication In addition to the above discussion on how ToM is used in LLM research, we have identified four common types of misuse when citing ToM-related papers. Although these misuses specifically occur in the context of referencing ToM works, they also reflect broader issues in how current LLM research draws on findings from psychology research.

- **Conceptual overgeneralization and misclassification** One of the most common misuses in LLM research when citing ToM-related papers is the overgeneralization and misclassification of the ToM concept. Researchers often reference psychology research on ToM without sufficiently attending to the original studies' specific research designs, populations, and experimental conclusions. As a result, ToM is frequently treated as a catch-all label, with key distinctions between task types and cognitive processes overlooked. Ideally, when citing a paper, researchers should clearly convey the study's core findings and scope of applicability, then thoughtfully relate these to the observed behaviors or capacities of LLMs. However, many LLM studies adopt ToM-related terminology, such as belief reasoning, perspective-taking, or false belief tasks, without a deep understanding of the cited studies. A common example of such conceptual overgeneralization is the failure to distinguish between different levels of mental-state reasoning tasks, such as first-order and second-order ToM, referring to them broadly as "ToM tasks" without clarifying their differing cognitive demands and underlying psychological mechanisms. For instance, some studies treat the Sally-Anne task (a first-order false belief task) and the Ice Cream Van task (a second-order false belief task) as equivalent tests of ToM when evaluating whether LLMs possess ToM abilities. Yet in psychology, these tasks are clearly differentiated: the former requires understanding that someone can hold a false belief about the world, while the latter involves reasoning about someone's belief about another person's belief (i.e., a recursive representation of mental states). LLM researchers need to reflect more deeply on why a particular paper is being cited over others, and what specific conclusions or experimental designs from that study meaningfully contribute to LLM research.

  In addition, there is a tendency in some LLM research to mistakenly categorize psychology studies that were not originally intended to explore ToM as supporting evidence for ToM. For example, studies focusing on emotion recognition, social attention, or attribution mechanisms are sometimes included in discussions of ToM. Although these studies are indeed related to social cognition, they do not strictly fall within the core components of ToM, especially when they do not involve recursive belief construction or the understanding of mental states. For example, emotion recognition and belief attribution are distinct psychological processes. The former relies more on perception, while the latter involves recursive modeling of mental states. This practice may stem from a vague understanding of what constitutes ToM, or from a selective interpretation of psychology research when constructing arguments, thereby weakening the conceptual rigor of ToM in interdisciplinary research.
- **Partial or incomplete citation** Another common misuse is that when citing psychology research on ToM, researchers often only select a few well-known studies while overlooking other equally important but less "representative" work in the field. We understand that citing well-known papers helps strengthen the credibility and authority of a paper, especially in literature reviews or theoretical frameworks, where referencing widely recognized classic studies can build a solid academic foundation. However, this practice can also lead to a narrow research perspective and neglect the diversity of findings within the field. In fact, some less "representative" studies, although less well-known, still offer valuable contributions that complement, challenge, or deepen mainstream views through their methods, samples, or conclusions. These can be especially useful as references for certain types of LLM research on ToM. For example, most LLM studies referencing psychology research on ToM often cite Premack and Woodruff (1978) work on whether chimpanzees possess ToM, or Baron-Cohen et al. (1985) study on ToM in children with autism. Although these studies are undoubtedly milestones in ToM research history, they may not always be the most suitable for exploring ToM in the context of LLMs. Specifically, some LLM-related studies want to examine how reasoning abilities evolve across different training stages or

attempt to break down the model's "ToM-like" capabilities into multiple levels for analysis. In this context, the study by Wellman and Liu (2004) on the developmental progression of ToM may offer more relevant insights. Their research designed a set of hierarchically structured tasks to demonstrate that children's ToM abilities develop gradually through several stages. This phased, progressive perspective aligns more closely with how LLMs' capabilities are described and offers a more structured theoretical framework for refining the analysis of "ToM-like" abilities.

Moreover, this emphasis on well-known research has led LLM research to primarily draw on a limited number of widely cited experimental findings, while overlooking many studies that also hold significant theoretical value and experimental insight. For example, Apperly et al. (2006) found that even adults do not automatically engage ToM abilities in certain contexts; instead, they rely on cognitive control resources to perform reasoning tasks about mental states. This finding can be important for understanding the conditions under which LLMs might exhibit an active use of ToM-like abilities. Similarly, Onishi and Baillargeon (2005), using the violation of expectation paradigm with infants, demonstrated that sensitivity to others' belief states may emerge at an earlier developmental stage than previously thought. This provides empirical support for the concept of "early ToM" and offers valuable clues for exploring whether LLMs might develop some form of ToM processing during pretraining. Although these less frequently cited studies are not as widely known as classic experiments by Baron-Cohen et al. (1985) or Premack and Woodruff (1978), they offer unique value to current LLM research in terms of methodology, theoretical perspective, and task design. Continued neglect of such studies in LLM research risks missing important opportunities to deepen our understanding of ToM-like capabilities in LLM research.

- **Misinterpretation or misrepresentation of findings** Misinterpretation of cited psychology research is also a common issue in LLM research. This may have led some LLM studies to cite inappropriate papers to support certain arguments. This problem may stem from a lack of disciplinary sensitivity among LLM researchers when dealing with interdisciplinary literature, which may result in the use of studies that appear relevant on the surface but fail to provide adequate support. For example, some LLM research cites psychology studies on the biological mechanisms of ToM to support discussions about ToM performance at the social level. This conflates theoretical frameworks and research goals across different levels. The activation of a specific brain region may indicate how a certain function is recruited, but it cannot directly show the individual's behavioral strategies or interaction styles in specific social contexts. Some researchers may use the superficial criterion of "this article studied ToM" to include it as supporting material, thereby masking logical leaps in their argumentation.

  In addition, some LLM research selectively emphasizes positive findings while overlooking the limitations of the psychology research they cite. For example, researchers may highlight only the conclusive statements from cited research, while ignoring important caveats noted by the original authors, such as experimental boundaries, sample limitations, or theoretical controversies. In some cases, even papers that have been widely questioned (have controversial opinions) are cited without any clarification. This kind of selective referencing can lead to a biased evidentiary base, thereby weakening the rigor of arguments. For example, some papers refer to the two-systems account of ToM and observational studies on children's gaze behaviors to support the idea that LLMs may possess implicit ToM abilities, yet they fail to adequately address the ongoing debates and limitations within these studies. Notably, there is no scholarly consensus on whether children's gaze-shifting behavior truly indicates an implicit ToM. Some researchers suggest that such behaviors may stem from low-level attentional preferences rather than genuine mental state attribution. We recommend that LLM researchers adopt a critical mindset when referencing psychology research, presenting the theoretical context, methodological constraints, and academic debates of the cited studies in a balanced way to ensure accuracy and scientific rigor in their arguments.

- **Secondary citation errors** Another characteristic of how LLM research cites ToM papers is that researchers tend to rely more on secondary literature within the AI community when constructing ToM-related arguments, rather than directly consulting or citing primary literature from the field of psychology. In other words, many theoretical claims or experimental justifications concerning ToM are based on interpretations by other LLM/NLP researchers rather than on direct engagement with original psychology research. This phenomenon is not uncommon in scientific communication and within knowledge communities, and it is easy to understand why: researchers are more inclined to cite well-recognized work within their own field to enhance the

acceptability and persuasiveness of their arguments during submission, peer review, or academic evaluation. This "in-group citation preference" tends to create a closed citation loop. Early studies that introduced ToM into the LLM context established a framework of terminology and tasks, and subsequent research continues to build upon this framework, gradually forming an internally-reinforced citation network.

However, the cumulative bias inherent in secondary citations can further amplify the previously mentioned issues, such as conceptual overgeneralization, task confusion, narrow literature selection, and misinterpretation. If one psychology research is misunderstood during its initial citation, subsequent literature that continues to cite this interpretation without verification can lead to a "consensus of misreading." More critically, the nuanced descriptions, methodological complexities, and theoretical debates surrounding ToM in original psychology research are often compressed, oversimplified, or even omitted in secondary citations. For example, some papers may not directly engage with the original studies by Baron-Cohen et al. (1985) on autism and ToM, but instead rely on brief summaries from other LLM research, which can risk overlooking essential elements such as experimental controls and sample differences. We recommend that LLM researchers trace psychology theories or experiments back to their original sources, rather than relying solely on interpretative summaries from the NLP/LLM community. Only by returning to the original citations can one clarify the theoretical context, research intent, and methodological limitations, thereby ensuring accuracy in understanding and rigor in application.

6 Discussion

6.1 Summary of key findings

Here, we summarize the answers to the three research questions posed at the end of §1.

RQ1: How is psychology research integrated into LLM research?

A1: Since 2023, an increasing number of LLM research papers have cited psychology papers, indicating a growing interest among researchers in insights from psychology. So far, psychology has been broadly integrated into LLM research, with *Neural Mechanisms* and *Psychometrics & JDM* being the most prominent psychology topics. LLM research topics demonstrate distinct referencing preferences for different areas of psychology. For example, *Educational Application* and *Advanced Reasoning* clearly favor psychology papers from the *Education* and *Neural Mechanisms* clusters, respectively, whereas *Model Adaptation & Efficiency* and *Social Intelligence* draw on a much broader range of psychology topics. These citation patterns result from the interplay between the nature of research topics in psychology and LLM research, as well as the characteristics of LLMs, such as being fast-updating, data-intensive, and black-box. More detailed discussions can be found in §5.1.

RQ2: Which psychology theories/frameworks are most commonly used, and which remain underexplored in LLM research?

A2: The top 10 psychology theories and frameworks most frequently cited by the surveyed LLM research papers are Dual-Process Theories (434 citations), Theory of Mind (309 citations), Heuristics and Biases Program (210 citations), Executive Functions (197 citations), Connectionism vs. Symbolism (190 citations), Theory-Theory (151 citations), Classical Test Theory (145 citations), Usage-based Models of Language (128 citations), Mental Simulation Theory (113 citations), and Schema Theory (109 citations). The application of these theories and frameworks reveals that the LLM research paradigm is becoming increasingly pluralistic under the influence of psychology. This paradigm begins as performance-driven and model-centric. As more psychology theories and frameworks are incorporated to guide experiments on LLMs, it gradually adopts a theory-driven and data-centric, i.e., empirical, approach. However, despite the many theories and frameworks within each psychology topic, as elaborated on in §5.2, only some have been engaged by LLM research, leaving ample room for further exploration.

RQ3: How is psychology research operationalized and interpreted in the context of LLM research?

A3: A psychology theory/framework has multiple facets that can be studied from different perspectives in psychology research, leading to a variety of applications in LLM research that cite them. A case study on Theory of Mind is presented in §5.3 to exemplify this diversity. Across the different methods of operationalization and interpretation, there are four types of common misapplications: 1)

Conceptual Overgeneralization and Misclassification, where LLM researchers cite related psychology research clarifying its primary design, target population, or key conclusions; 2) *Partial or Incomplete Citation*, where LLM researchers rely on a few popular papers about their intended psychology theories and frameworks, overlooking other, potentially more relevant, but less well-known research; 3) *Misinterpretation or Misrepresentation of Findings*, where LLM researchers cite inappropriate psychology papers to support their arguments or overly emphasize partial findings from the cited papers, despite only topical relevance; and 4) *Secondary Citation Errors*, where LLM researchers prefer citing influential LLM research papers that engage with the intended psychology theories and frameworks over the original psychology research itself. All these misuses, while they may lead to surprising findings, could compromise the validity and accuracy of insights drawn from psychology.

6.2 Theoretical and methodological reflections

Although the intersection of AI research and psychology is advancing rapidly, productive interdisciplinary integration continues to face theoretical and methodological challenges.

First, on the theoretical level, current AI research often tends to instrumentalize psychology theories, simplifying complex ideas into quantifiable conceptual labels. Although this simplification lowers the threshold for applying theories, it can also obscure the deeper structures of the original frameworks. For example, the use of ToM in AI research often treats mental states like beliefs and desires as static data points. In contrast, within psychology research, ToM is understood as a dynamic, context-dependent ability, typically characterized by uncertainty, ambiguity, and developmental variability. Behind this simplification lies a fundamental difference between the goals of psychology theories and the goals of AI researchers when applying these theories. Psychology aims to explain the internal mechanisms of human behavior and mental activity, emphasizing the complexity of processes and the importance of social and cultural contexts. In contrast, AI researchers tend to focus more on functional reconstruction and engineering implementation, often transforming psychology concepts into operational and computable model parameters. The risk of such oversimplification in AI can lead to misleading conclusions about the capabilities of AI systems, such as overestimating their understanding of human behavior or wrongly attributing human-like intentionality to models that merely simulate behavioral patterns. By failing to account for the fluid, interpretive nature of social reasoning, AI models risk reinforcing shallow imitations of human cognition, which may perform well in isolated tasks but lack the flexibility and depth needed for genuinely adaptive or ethical interaction. To avoid conceptual drift, AI researchers referencing psychology theories should pay close attention to their original context, like core assumptions and theoretical boundaries, rather than merely adopting surface-level terminology.

Additionally, psychology theories are often used as tools to explain model behavior, but such post hoc explanations often lack predictive power and systematic structure. One example is to use attention in cognitive psychology to explain the behavior of neural models like Transformers. This explanation is derived after observing the model's output and does not offer predictive insight into how the model will behave in unseen scenarios. The psychology concept of attention involves complex processes, whereas the attention mechanism in models is a deterministic computation of similarity scores. Without a rigorous mapping between theoretical constructs and model components, these psychology references risk becoming superficial narratives—appealing and intuitive, but ultimately can not guide future model development or evaluation. Interdisciplinary research should place greater emphasis on theoretical modeling in the early stages, encouraging the integration of clear psychological hypotheses during the experimental design phase, rather than retrofitting existing theories only at the analysis stage.

At the methodological level, the operationalization of psychological concepts in AI research also faces challenges. Many studies attempt to simulate psychological phenomena by constructing proxy tasks; however, these tasks are often unvalidated and lack construct validity. We argue that AI research should draw on psychology's strong emphasis on measurement validity and experimental control by incorporating more systematic psychological methods into task design and data interpretation. For example, collaborating with psychologists to design experiments, using standardized scales, and reporting the psychological validity of tasks are all feasible strategies for improving methodological quality.

It is worth noting that the field of HCI has long been a prime example of interdisciplinary research. This field not only values theory-driven research design but also emphasizes methodological diversity

and rigor. Drawing from multiple disciplines, HCI has successfully embedded different knowledge into multiple stages of the research process, such as problem definition, system design, user modeling, and evaluation. This systematic approach offers valuable insights for AI research.

6.3 Toward more responsible interdisciplinary practice

To build a stronger and more sustainable bridge between AI and psychology research, we advocate for more responsible interdisciplinary practices. This is not just about upholding ethical standards; it is also about ensuring scientific rigor. With that in mind, we offer a few recommendations aimed at promoting clearer standards, more consistent methods, and deeper collaboration across disciplines.

Theoretical accountability. When drawing on psychology theories, researchers should clearly explain where these theories come from, what assumptions they rest on, and how far they can reasonably be applied to avoid misinterpretation or conceptual reconstruction. Furthermore, competing theoretical perspectives should be adequately addressed, with explicit justification for the chosen framework and acknowledgment of its limitations. Such theoretical accountability not only strengthens research transparency but also provides a thoughtful re-evaluation of the extent to which psychology theories can be meaningfully applied in AI research.

Construct operationalization. Interdisciplinary research works best when there is a clear and defensible connection between psychological concepts and the technical tasks. We suggest starting with standardized, widely accepted measurement tools from psychology whenever possible. If you need to design custom tasks, it is important to clearly explain how those tasks reflect the underlying psychological construct and to back that explanation with theory. When appropriate, multiple forms of measurement (e.g., behavioral data, linguistic outputs, and subjective ratings) should be employed to ensure a comprehensive evaluation of the construct.

Collaborative parity. Interdisciplinary collaboration should not be about one field just doing what another wants. Instead, it really needs to be built on a foundation of mutual respect, where different disciplines are truly co-creating something new. To make this happen, we think it is important to encourage things like joint authorship across disciplines, working together from the very beginning to formulate research questions, and making sure we weave in a variety of analytical perspectives.

Open interdisciplinary infrastructure. We advocate for the development of open knowledge infrastructures that support interdisciplinary collaboration. Examples include reusable datasets on psychology constructs and measurement methods, case templates for interdisciplinary research, and cross-referencing maps. Such resources would lower the barriers to collaboration, enhance the quality of research, and foster the accumulation and transmission of knowledge across disciplinary communities.

In sum, responsible interdisciplinary practice is not a matter of occasional collaboration, but rather a sustainable and institutionalized research framework. By strengthening theoretical accountability, standardizing how we use constructs, and promoting equitable collaboration and shared infrastructure, we can achieve a deeper, more trusting integration of AI and psychology research.

6.4 Limitations and future directions

Although we aim to provide a thorough survey of how psychology research is cited and integrated in AI research, several limitations should be acknowledged, as they may affect the broader applicability and interpretive depth of our findings.

First, in terms of the time range, our analysis focuses on published CS research between December 2022 and March 2025, primarily reflecting the short-term impact and initial integration of psychology research into LLM research. Given that interdisciplinary influences often exhibit a time lag, the deeper transformation of psychology theories, methodological integration, and practical impact may still be in the early stages. As a result, our findings may underestimate the long-term knowledge diffusion and paradigm-shifting role of psychology in LLM research.

Secondly, in terms of the analyzed data, we primarily examined English-based papers from top AI conferences, which may have led to the omission of relevant work from other important fields.

Additionally, differences in database indexing mechanisms and citation formats may have caused some psychology sources to be overlooked, potentially resulting in a possible underrepresentation of psychology sources.

Purely exploring the integration of AI research and psychology research from the citation-based perspective might be a limitation. Future studies could extend this work by moving beyond citation-based analyses to explore the actual processes of interdisciplinary collaboration between AI and psychology. This may include examining how knowledge is negotiated across disciplinary boundaries, how research teams are structured, and which collaboration mechanisms are most effective. Such qualitative insights would complement the current study's bibliometric approach and provide a more comprehensive understanding of how these fields interact in practice.

7 Conclusion

This work contributes to a growing science of science approach to understanding how interdisciplinary knowledge circulates, mutates, and influences AI development. By identifying the domains and dynamics of psychology research influence in LLM research, we aim to provide not only a descriptive map but also a normative guide: showing how psychology research is most productively integrated, where misuse arises, and how better practices can be cultivated. As AI systems become increasingly embedded in the fabric of society, the importance of methodological pluralism, conceptual clarity, and cross-disciplinary rigor will only grow. Psychology has helped us understand human intelligence; with care and collaboration, it may also help us build AI more wisely.

Acknowledgments and Disclosure of Funding

This research project is partially sponsored by the Microsoft Accelerate Foundation Models Research (AFMR) grant program.

References

Abdulhai, M., Serapio-García, G., Crepy, C., Valter, D., Canny, J., Jaques, N., 2024. Moral foundations of large language models, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 17737–17752. URL: https://aclanthology.org/2024.emnlp-main.982/, doi:10.18653/v1/2024.emnlp-main.982.

Alabdulmohsin, I.M., Neyshabur, B., Zhai, X., 2022. Revisiting neural scaling laws in language and vision, in: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 22300–22312. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/8c22e5e918198702765ecff4b20d0a90-Paper-Conference.pdf.

Alabed, A., Javornik, A., Gregory-Smith, D., 2022. Ai anthropomorphism and its effect on users' self-congruence and self–ai integration: A theoretical framework and research agenda. Technological Forecasting and Social Change 182, 121786. URL: https://www.sciencedirect.com/science/article/pii/S0040162522003109, doi:https://doi.org/10.1016/j.techfore.2022.121786.

Alam, M.T., Nguyen, L., Bhusal, D., Rastogi, N., 2025. Ctibench: a benchmark for evaluating llms in cyber threat intelligence, in: Proceedings of the 38th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA.

Alizadeh, K., Mirzadeh, S.I., Belenko, D., Khatamifard, S., Cho, M., Del Mundo, C.C., Rastegari, M., Farajtabar, M., 2024. LLM in a flash: Efficient large language model inference with limited memory, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 12562–12584. URL: https://aclanthology.org/2024.acl-long.678/, doi:10.18653/v1/2024.acl-long.678.

American Psychological Association, n.d. Apa divisions. https://www.apa.org/about/division. Accessed: 2025-06-07.

An, H., Acquaye, C., Wang, C., Li, Z., Rudinger, R., 2024. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender?, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 386–397. URL: https://aclanthology.org/2024.acl-short.37/, doi:10.18653/v1/2024.acl-short.37.

Anderson, J.R., 2013. The adaptive character of thought. Psychology Press.

Anthropic, 2024. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. Accessed: 2025-07-02.

Apperly, I.A., Riggs, K.J., Simpson, A., Chiavarino, C., Samson, D., 2006. Is belief reasoning automatic? Psychological Science 17, 841–844.

Plaza-del Arco, F.M., Curry, A.C., Paoli, S., Cercas Curry, A., Hovy, D., 2024. Divine LLaMAs: Bias, stereotypes, stigmatization, and emotion representation of religion in large language models, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA. pp. 4346–4366. URL: https://aclanthology.org/2024.findings-emnlp.251/, doi:10.18653/v1/2024.findings-emnlp.251.

Ashkinaze, J., Fry, E., Edara, N., Gilbert, E., Budak, C., 2025. Plurals: A system for guiding llms via simulated social ensembles, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. URL: https://doi.org/10.1145/3706598.3713675, doi:10.1145/3706598.3713675.

Association, A.P., 2013. Diagnostic and statistical manual of mental disorders: DSM-5, 5th ed. American Psychiatric Publishing, Inc., Arlington, VA, US. doi:https://doi.org/10.1176/appi.books.9780890425596.

Atkinson, R.C., Shiffrin, R.M., 1968. Human memory: A proposed system and its control processes, in: Psychology of learning and motivation. Elsevier. volume 2, pp. 89–195.

Baddeley, A., 2003. Working memory: looking back and looking forward. Nature reviews neuroscience 4, 829–839.

Baddeley, A., 2020. Working memory. Memory , 71–111.

Bakeman, R., Quera, V., 2011. Sequential analysis and observational methods for the behavioral sciences. Cambridge University Press.

Bar-Hillel, M., 1980. The base-rate fallacy in probability judgments. Acta Psychologica 44, 211–233.

Baron-Cohen, S., Leslie, A.M., Frith, U., 1985. Does the autistic child have a "theory of mind" ? Cognition 21, 37–46. URL: https://www.sciencedirect.com/science/article/pii/0010027785900228, doi:https://doi.org/10.1016/0010-0277(85)90022-8.

Beck, J.S., 2011. Cognitive behavior therapy: Basics and beyond (2nd ed). The Guilford Press, New York, US.

Beed, P.L., Hawkins, E.M., Roller, C.M., 1991. Moving learners toward independence: The power of scaffolded instruction. The Reading Teacher 44, 648–655. URL: http://www.jstor.org/stable/20200767.

Belinkov, Y., Glass, J., 2019. Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics 7, 49–72.

Benjamin Jr, L.T., 2023. A brief history of modern psychology. John Wiley & Sons.

Berko, J., 1958. The child's learning of english morphology. Word 14, 150–177.

Bernabei, M., Colabianchi, S., Falegnami, A., Costantino, F., 2023. Students' use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances. Computers and Education: Artificial Intelligence 5, 100172. URL: https://www.sciencedirect.com/science/article/pii/S2666920X23000516, doi:https://doi.org/10.1016/j.caeai.2023.100172.

Bertolazzi, L., Gatt, A., Bernardi, R., 2024. A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 13882–13905. URL: https://aclanthology.org/2024.emnlp-main.769/, doi:10.18653/v1/2024.emnlp-main.769.

Bianchi, F., Chia, P.J., Yuksekgonul, M., Tagliabue, J., Jurafsky, D., Zou, J., 2024. How well can llms negotiate? negotiationarena platform and analysis, in: Proceedings of the 41st International Conference on Machine Learning, JMLR.org.

Blandford, A., 2019. Hci for health and wellbeing: Challenges and opportunities. International Journal of Human-Computer Studies 131, 41–51. URL: https://www.sciencedirect.com/science/article/pii/S1071581919300771, doi:https://doi.org/10.1016/j.ijhcs.2019.06.007. 50 years of the International Journal of Human-Computer Studies. Reflections on the past, present and future of human-centred technologies.

Bloom, B.S., 1956. Taxonomy of educational objectives: The classification of educational goals. Longman Group.

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al., 2022. On the opportunities and risks of foundation models. arXiv:2108.07258.

Bookheimer, S., 2002. Functional mri of language: new approaches to understanding the cortical organization of semantic processing. Annual review of neuroscience 25, 151–188.

Borges, B., Tandon, N., Käser, T., Bosselut, A., 2024. Let me teach you: Pedagogical foundations of feedback for language models, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 12082–12104. URL: https://aclanthology.org/2024.emnlp-main.674/, doi:10.18653/v1/2024.emnlp-main.674.

Boschi, V., Catricala, E., Consonni, M., Chesi, C., Moro, A., Cappa, S.F., 2017. Connected speech in neurodegenerative language disorders: a review. Frontiers in psychology 8, 269.

Broadbent, D., 1958. Chapter 2 - selective listening to speech, in: Broadbent, D. (Ed.), Perception and Communication. Pergamon, pp. 11–35. URL: https://www.sciencedirect.com/science/article/pii/B9781483200798500049, doi:https://doi.org/10.1016/B978-1-4832-0079-8.50004-9.

Brown, J.D., 1986. Evaluations of self and others: Self-enhancement biases in social judgments. Social cognition 4, 353–376.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Bunge, M., 2017. Philosophy of science: Volume 1, from problem to theory. Routledge.

Cao, B., Ren, M., Lin, H., Han, X., Zhang, F., Zhan, J., Sun, L., 2024. StructEval: Deepen and broaden large language model assessment via structured evaluation, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand. pp. 5300–5318. URL: https://aclanthology.org/2024.findings-acl.314/, doi:10.18653/v1/2024.findings-acl.314.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P., Sun, L., 2025. A survey of ai-generated content (aigc). ACM Comput. Surv. 57. URL: https://doi.org/10.1145/3704262, doi:10.1145/3704262.

Chakrabarty, T., Saakyan, A., Winn, O., Panagopoulou, A., Yang, Y., Apidianaki, M., Muresan, S., 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada. pp. 7370–7388. URL: https://aclanthology.org/2023.findings-acl.465/, doi:10.18653/v1/2023.findings-acl.465.

Chan, C., Jiayang, C., Yim, Y., Deng, Z., Fan, W., Li, H., Liu, X., Zhang, H., Wang, W., Song, Y., 2024. NegotiationToM: A benchmark for stress-testing machine theory of mind on negotiation surrounding, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA. pp. 4211–4241. URL: https://aclanthology.org/2024.findings-emnlp.244/, doi:10.18653/v1/2024.findings-emnlp.244.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., Xie, X., 2024. A survey on evaluation of large language models. ACM transactions on intelligent systems and technology 15, 1–45. URL: https://doi.org/10.1145/3641289, doi:10.1145/3641289.

Chen, H., Dou, Z., Mao, K., Liu, J., Zhao, Z., 2024a. Generalizing conversational dense retrieval via LLM-cognition data augmentation, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 2700–2718. URL: https://aclanthology.org/2024.acl-long.149/, doi:10.18653/v1/2024.acl-long.149.

Chen, Y., Xing, X., Lin, J., Zheng, H., Wang, Z., Liu, Q., Xu, X., 2023. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 1170–1183. URL: https://aclanthology.org/2023.findings-emnlp.83/, doi:10.18653/v1/2023.findings-emnlp.83.

Chen, Z., Li, D., Zhao, X., Hu, B., Zhang, M., 2024b. Temporal knowledge question answering via abstract reasoning induction, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 4872–4889. URL: https://aclanthology.org/2024.acl-long.267/, doi:10.18653/v1/2024.acl-long.267.

Cherry, E.C., 1953. Some experiments on the recognition of speech, with one and with two ears. Journal of the Acoustical Society of America 25, 975–979. URL: https://hdl.handle.net/11858/00-001M-0000-002A-F750-3, doi:10.1121/1.1907229.

Chomsky, N., 2002. Syntactic structures. Mouton de Gruyter.

Citri, A., Malenka, R.C., 2008. Synaptic plasticity: Multiple forms, functions, and mechanisms. Neuropsychopharmacology 33, 18–41. URL: https://doi.org/10.1038/sj.npp.1301559, doi:10.1038/sj.npp.1301559.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D., 2020. SPECTER: Document-level representation learning using citation-informed transformers, in: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 2270–2282. URL: https://aclanthology.org/2020.acl-main.207/, doi:10.18653/v1/2020.acl-main.207.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46. URL: https://doi.org/10.1177/001316446002000104, doi:10.1177/001316446002000104, arXiv:https://doi.org/10.1177/001316446002000104.

Cohen, J., 1994. The earth is round (p<. 05). American psychologist 49, 997.

Colman, A., 2016. What is psychology? Routledge.

Coon, D., Mitterer, J.O., 2013. Introduction to psychology: Gateways to mind and behavior. Wadsworth Cengage Learning.

Cosmides, L., Tooby, J., 1996. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. cognition 58, 1–73.

Crawford, K., 2021. The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

Cuddy, A.J., Fiske, S.T., Kwan, V.S., Glick, P., Demoulin, S., Leyens, J.P., Bond, M.H., Croizet, J.C., Ellemers, N., Sleebos, E., et al., 2009. Stereotype content model across cultures: Towards universal similarities and some differences. British journal of social psychology 48, 1–33.

Cuijpers, P., Li, J., Hofmann, S.G., Andersson, G., 2010. Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: a meta-analysis. Clinical psychology review 30, 768–778.

Cuskley, C., Woods, R., Flaherty, M., 2024. The limitations of large language models for understanding human language and cognition. Open Mind 8, 1058–1083. URL: https://doi.org/10.1162/opmi_a_00160, doi:10.1162/opmi_a_00160, arXiv:https://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi_a_00160/2468254/opmi_a_00160.pdf.

Dagan, G., Synnaeve, G., Rozière, B., 2024. Getting the most out of your tokenizer for pre-training and domain adaptation, in: Proceedings of the 41st International Conference on Machine Learning, JMLR.org.

Daheim, N., Macina, J., Kapur, M., Gurevych, I., Sachan, M., 2024. Stepwise verification and remediation of student reasoning errors with large language model tutors, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 8386–8411. URL: https://aclanthology.org/2024.emnlp-main.478/, doi:10.18653/v1/2024.emnlp-main.478.

Damasio, A.R., 2006. Descartes' error. Random House.

Deci, E.L., Ryan, R.M., 2013. Intrinsic motivation and self-determination in human behavior. Springer Science & Business Media.

DeFleur, M.L., 1964. Stigma: Notes on the management of spoiled identity. by erving goffman. englewood cliffs, new jersey: Prentice-hall, 1963. 147 pp. cloth, $4.50; paper, $1.95. Social Forces 43, 127–128. URL: https://doi.org/10.1093/sf/43.1.127, doi:10.1093/sf/43.1.127, arXiv:https://academic.oup.com/sf/article-pdf/43/1/127/6506421/43-1-127.pdf.

Desimone, R., Duncan, J., et al., 1995. Neural mechanisms of selective visual attention. Annual review of neuroscience 18, 193–222.

Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L., 2023. Qlora: efficient finetuning of quantized llms, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA.

Dewey, J., 1892. Psychology. American Book Company.

Diamond, A., 2013. Executive functions. Annual Review of Psychology 64, 135–168. doi:10.1146/annurev-psych-113011-143750.

Didolkar, A.R., Goyal, A., Ke, N.R., Guo, S., Valko, M., Lillicrap, T.P., Rezende, D.J., Bengio, Y., Mozer, M.C., Arora, S., 2024. Metacognitive capabilities of LLMs: An exploration in mathematical problem solving, in: The Thirty-eighth Annual Conference on Neural Information Processing Systems. URL: https://openreview.net/forum?id=D19UyP4HYk.

Diener, E., Emmons, R.A., Larsen, R.J., Griffin, S., 1985. The satisfaction with life scale. Journal of personality assessment 49, 71–75.

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D.w., Oishi, S., Biswas-Diener, R., 2010. New well-being measures: Short scales to assess flourishing and positive and negative feelings. Social indicators research 97, 143–156.

de Dios-Flores, I., Garcia Amboage, J., Garcia, M., 2023. Dependency resolution at the syntax-semantics interface: psycholinguistic and computational insights on control dependencies, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada. pp. 203–222. URL: https://aclanthology.org/2023.acl-long.12/, doi:10.18653/v1/2023.acl-long.12.

Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., Sui, Z., 2024. A survey on in-context learning, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 1107–1128. URL: https://aclanthology.org/2024.emnlp-main.64/, doi:10.18653/v1/2024.emnlp-main.64.

Dong, W., Zhao, Y., Sun, Z., Liu, Y., Peng, Z., Zheng, J., Zhang, Z., Zhang, Z., Wu, J., Wang, R., Xu, S., Huang, X., He, X., 2025. Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications. URL: https://arxiv.org/abs/2505.00049, arXiv:2505.00049.

Dong, Y., Wang, Z., Sreedhar, M., Wu, X., Kuchaiev, O., 2023. SteerLM: Attribute conditioned SFT as an (user-steerable) alternative to RLHF, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 11275–11288. URL: https://aclanthology.org/2023.findings-emnlp.754/, doi:10.18653/v1/2023.findings-emnlp.754.

Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I., 2024. Improving factuality and reasoning in language models through multiagent debate, in: Proceedings of the 41st International Conference on Machine Learning, JMLR.org.

Dziri, N., Lu, X., Sclar, M., Li, X.L., Jiang, L., Lin, B.Y., Welleck, S., West, P., Bhagavatula, C., Bras, R.L., Hwang, J.D., Sanyal, S., Ren, X., Ettinger, A., Harchaoui, Z., Choi, Y., 2023. Faith and fate: Limits of transformers on compositionality, in: Thirty-seventh Conference on Neural Information Processing Systems. URL: https://openreview.net/forum?id=Fkckkr3ya8.

Echterhoff, J.M., Liu, Y., Alessa, A., McAuley, J., He, Z., 2024. Cognitive bias in decision-making with LLMs, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA. pp. 12640–12653. URL: https://aclanthology.org/2024.findings-emnlp.739/, doi:10.18653/v1/2024.findings-emnlp.739.

Ehrlich, S.F., Rayner, K., 1981. Contextual effects on word perception and eye movements during reading. Journal of verbal learning and verbal behavior 20, 641–655.

Ekman, P., 1992. An argument for basic emotions. Cognition and Emotion 6, 169–200. doi:10.1080/02699939208411068.

Elliott, R., Bohart, A.C., Watson, J.C., Murphy, D., 2018. Therapist empathy and client outcome: An updated meta-analysis. Psychotherapy (Chicago, Ill.) 55, 399–410. doi:10.1037/pst0000175. meta-analysis.

Fan, J., Saaty, M., Mccrickard, D.S., 2024a. Education in hci outdoors: A diary study approach, in: Proceedings of the 6th Annual Symposium on HCI Education, Association for Computing Machinery, New York, NY, USA. URL: https://doi.org/10.1145/3658619.3658621, doi:10.1145/3658619.3658621.

Fan, S., Pagliardini, M., Jaggi, M., 2024b. Doge: domain reweighting with generalization estimation, in: Proceedings of the 41st International Conference on Machine Learning, JMLR.org.

Faules, D.F., Alexander, D.C., 1978. Communication and social behavior : a symbolic interaction perspective. Reading (Mass.) : Addison-Wesley.

Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Huang, A., Zhang, S., Chen, K., Yin, Z., Shen, Z., Ge, J., Ng, V., 2024. LawBench: Benchmarking legal knowledge of large language models, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 7933–7962. URL: https://aclanthology.org/2024.emnlp-main.452/, doi:10.18653/v1/2024.emnlp-main.452.

Feng, S., Sorensen, T., Liu, Y., Fisher, J., Park, C.Y., Choi, Y., Tsvetkov, Y., 2024. Modular pluralism: Pluralistic alignment via multi-LLM collaboration, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 4151–4171. URL: https://aclanthology.org/2024.emnlp-main.240/, doi:10.18653/v1/2024.emnlp-main.240.

Feuerriegel, S., Hartmann, J., Janiesch, C., Zschech, P., 2024. Generative ai. Business & Information Systems Engineering 66, 111–126. URL: https://doi.org/10.1007/s12599-023-00834-7, doi:10.1007/s12599-023-00834-7.

Fitzpatrick, K.K., Darcy, A., Vierhile, M., 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. JMIR mental health 4, e7785.

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin 76, 378–382. doi:10.1037/h0031619.

Floridi, L., 2023. The ethics of artificial intelligence: Principles, challenges, and opportunities. Oxford University Press.

Fodor, J.A., Pylyshyn, Z.W., 1988. Connectionism and cognitive architecture: A critical analysis. Cognition 28, 3–71. URL: https://www.sciencedirect.com/science/article/pii/0010027788900315, doi:https://doi.org/10.1016/0010-0277(88)90031-5.

Forde, J.Z., Zhang, R., Sutawika, L., Aji, A.F., Cahyawijaya, S., Winata, G.I., Wu, M., Eickhoff, C., Biderman, S., Pavlick, E., 2024. Re-evaluating evaluation for multilingual summarization, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 19476–19493. URL: https://aclanthology.org/2024.emnlp-main.1085/, doi:10.18653/v1/2024.emnlp-main.1085.

Fowler Jr, F.J., 2013. Survey research methods. Sage publications.

French, R.M., 1999. Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences 3, 128–135. doi:10.1016/s1364-6613(99)01294-2. pMID: 10322466.

Gabriel, S., Puri, I., Xu, X., Malgaroli, M., Ghassemi, M., 2024. Can AI relate: Testing large language model response for mental health support, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA. pp. 2206–2221. URL: https://aclanthology.org/2024.findings-emnlp.120/, doi:10.18653/v1/2024.findings-emnlp.120.

Gao, J., Gebreegziabher, S.A., Choo, K.T.W., Li, T.J.J., Perrault, S.T., Malone, T.W., 2024a. A taxonomy for human-llm interaction modes: An initial exploration, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, pp. 1–11.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., Wang, H., 2024b. Retrieval-augmented generation for large language models: A survey. URL: https://arxiv.org/abs/2312.10997, arXiv:2312.10997.

Garner, R., 1987. Metacognition and reading comprehension. Ablex Publishing.

Gawronski, B., Creighton, L.A., 2013. The Oxford handbook of social cognition. Oxford University Press. chapter Dual process theories. pp. 282–312.

Gemini, Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., Firat, O., Molloy, J., et al., 2024. Gemini: A family of highly capable multimodal models. URL: https://arxiv.org/abs/2312.11805, arXiv:2312.11805.

Gigerenzer, G., 1991. From tools to theories: A heuristic of discovery in cognitive psychology. Psychological review 98, 254.

Glover, G.H., 2011. Overview of functional magnetic resonance imaging. Neurosurgery Clinics of North America 22, 133.

Goldstein, A., Havin, M., Reichart, R., Goldstein, A., 2023. Decoding stumpers: Large language models vs. human problem-solvers, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 11644–11653. URL: https://aclanthology.org/2023.findings-emnlp.779/, doi:10.18653/v1/2023.findings-emnlp.779.

Gómez-Rodríguez, C., Williams, P., 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 14504–14528. URL: https://aclanthology.org/2023.findings-emnlp.966/, doi:10.18653/v1/2023.findings-emnlp.966.

Goodman, N.D., Frank, M.C., 2016. Pragmatic language interpretation as probabilistic inference. Trends in cognitive sciences 20, 818–829.

Goodwin, C.J., 2015. A history of modern psychology. John Wiley & Sons.

Gorno-Tempini, M.L., Hillis, A.E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S.F., Ogar, J.M., Rohrer, J.D., Black, S., Boeve, B.F., et al., 2011. Classification of primary progressive aphasia and its variants. Neurology 76, 1006–1014.

Gou, Z., Shao, Z., Gong, Y., yelong shen, Yang, Y., Duan, N., Chen, W., 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing, in: The Twelfth International Conference on Learning Representations. URL: https://openreview.net/forum?id=Sx038qxjek.

Graesser, A.C., Nakamura, G.V., 1982. The impact of a schema on comprehension and memory, Academic Press. volume 16 of *Psychology of Learning and Motivation*, pp. 59–109. URL: https://www.sciencedirect.com/science/article/pii/S0079742108605472, doi:https://doi.org/10.1016/S0079-7421(08)60547-2.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S.P., Ditto, P.H., 2013. Moral foundations theory: The pragmatic validity of moral pluralism, in: Advances in experimental social psychology. Elsevier. volume 47, pp. 55–130.

Graham, S., Harris, K.R., 1993. Self-regulated strategy development: Helping students with learning problems develop as writers. The Elementary School Journal 94, 169–181. doi:https://doi.org/10.1086/461758.

Greimel, K.V., Kröner-Herwig, B., 2011. Cognitive behavioral treatment (cbt). Textbook of tinnitus , 557–561.

Guest, O., Martin, A.E., 2021. How computational modeling can force theory building in psychological science. Perspectives on Psychological Science 16, 789–802.

Gui, G., Toubia, O., 2023. The challenge of using llms to simulate human behavior: A causal inference perspective. SSRN Electronic Journal URL: http://dx.doi.org/10.2139/ssrn.4650172, doi:10.2139/ssrn.4650172.

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O., Zhang, X., 2024. Large language model based multi-agents: A survey of progress and challenges, in: Larson, K. (Ed.), Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, International Joint Conferences on Artificial Intelligence Organization. pp. 8048–8057. URL: https://doi.org/10.24963/ijcai.2024/890, doi:10.24963/ijcai.2024/890. survey Track.

Gupta, D., Li, B., 2024. A training data recipe to accelerate a* search with language models, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA. pp. 6681–6695. URL: https://aclanthology.org/2024.findings-emnlp.391/, doi:10.18653/v1/2024.findings-emnlp.391.

Hamilton, K., Shih, S.I., Mohammed, S., 2016. The development and validation of the rational and intuitive decision styles scale. Journal of personality assessment 98, 523–535.

Han, C., Ji, H., 2025. Computation mechanism behind llm position generalization. URL: https://arxiv.org/abs/2503.13305, arXiv:2503.13305.

Hao, S., Gu, Y., Ma, H., Hong, J.J., Wang, Z., Wang, D.Z., Hu, Z., 2023. Reasoning with language model is planning with world model, in: The 2023 Conference on Empirical Methods in Natural Language Processing. URL: https://openreview.net/forum?id=VTWWvYtF1R.

Hardt, D., 2023. Ellipsis-dependent reasoning: a new challenge for large language models, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Toronto, Canada. pp. 39–47. URL: https://aclanthology.org/2023.acl-short.4/, doi:10.18653/v1/2023.acl-short.4.

Hartigan, J.A., Wong, M.A., 1979. Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28, 100–108. URL: http://www.jstor.org/stable/2346830.

Hebb, D.O., 1949. The organization of behavior; a neuropsychological theory. Wiley.

Hegarty, P., Ansara, Y.G., Barker, M.J., 2018. Nonbinary gender identities. Gender, sex, and sexualities: Psychological perspectives , 53–76.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J., 2021. Measuring massive multitask language understanding. Proceedings of the International Conference on Learning Representations (ICLR) .

Henrich, J., Heine, S.J., Norenzayan, A., 2010. The weirdest people in the world? Behavioral and brain sciences 33, 61–83.

Hoffman, M.L., 1996. Empathy and moral development. The annual report of educational psychology in Japan 35, 157–162.

Hornsby, A.N., Love, B.C., 2020. How decisions and the desire for coherency shape subjective preferences over time. Cognition 200, 104244.

Horton, J.J., 2023. Large language models as simulated economic agents: What can we learn from homo silicus? URL: https://arxiv.org/abs/2301.07543, arXiv:2301.07543.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2022. LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

Hu, T., Collier, N., 2024. Quantifying the persona effect in LLM simulations, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 10289–10307. URL: https://aclanthology.org/2024.acl-long.554/, doi:10.18653/v1/2024.acl-long.554.

Huang, D., Dai, J., Weng, H., Wu, P., Qing, Y., Cui, H., Guo, Z., Zhang, J., 2024a. Effilearner: Enhancing efficiency of generated code via self-optimization, in: The Thirty-eighth Annual Conference on Neural Information Processing Systems. URL: https://openreview.net/forum?id=KhwOuB0fs9.

Huang, J., Chang, K.C.C., 2023. Towards reasoning in large language models: A survey, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada. pp. 1049–1065. URL: https://aclanthology.org/2023.findings-acl.67/, doi:10.18653/v1/2023.findings-acl.67.

Huang, X.A., La Malfa, E., Marro, S., Asperti, A., Cohn, A.G., Wooldridge, M.J., 2024b. A notion of complexity for theory of mind via discrete world models, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA. pp. 2964–2983. URL: https://aclanthology.org/2024.findings-emnlp.167/, doi:10.18653/v1/2024.findings-emnlp.167.

Hunston, S., 2006. Corpus linguistics. Encyclopedia of Language & Linguistics , 234–248doi:10.1016/B0-08-044854-2/00944-5.

Huutoniemi, K., 2010. Evaluating interdisciplinary research. The Oxford handbook of interdisciplinarity 10, 309–320.

Imani, S., Du, L., Shrivastava, H., 2023. MathPrompter: Mathematical reasoning using large language models, in: Sitaram, S., Beigman Klebanov, B., Williams, J.D. (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track), Association for Computational Linguistics, Toronto, Canada. pp. 37–42. URL: https://aclanthology.org/2023.acl-industry.4/, doi:10.18653/v1/2023.acl-industry.4.

Jain, A., Mao, J., Mohiuddin, K., 1996. Artificial neural networks: a tutorial. Computer 29, 31–44. doi:10.1109/2.485891.

James, W., 1892. Psychology. H. Holt.

Jang, M., Lukasiewicz, T., 2023. Consistency analysis of ChatGPT, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore. pp. 15970–15985. URL: https://aclanthology.org/2023.emnlp-main.991/, doi:10.18653/v1/2023.emnlp-main.991.

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Chen, B., Sun, R., Wang, Y., Yang, Y., 2023. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset, in: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track. URL: https://openreview.net/forum?id=g0QovXbFw3.

Jiang, B., Xie, Y., Hao, Z., Wang, X., Mallick, T., Su, W.J., Taylor, C.J., Roth, D., 2024. A peek into token bias: Large language models are not yet genuine reasoners, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 4722–4756. URL: https://aclanthology.org/2024.emnlp-main.272/, doi:10.18653/v1/2024.emnlp-main.272.

Jiang, H., Yi, X., Wei, Z., Xiao, Z., Wang, S., Xie, X., 2025. Raising the bar: Investigating the values of large language models via generative evolving testing, in: Proceedings of the 42st International Conference on Machine Learning, JMLR.org.

Jiayang, C., Qiu, L., Chan, T., Fang, T., Wang, W., Chan, C., Ru, D., Guo, Q., Zhang, H., Song, Y., Zhang, Y., Zhang, Z., 2023. StoryAnalogy: Deriving story-level analogies from large language models to unlock analogical understanding, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore. pp. 11518–11537. URL: https://aclanthology.org/2023.emnlp-main.706/, doi:10.18653/v1/2023.emnlp-main.706.

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J.Y., Shi, X., Chen, P.Y., Liang, Y., Li, Y.F., Pan, S., Wen, Q., 2024. Time-LLM: Time series forecasting by reprogramming large language models, in: International Conference on Learning Representations (ICLR).

John, O.P., Srivastava, S., 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives, in: Handbook of personality: Theory and research, 2nd ed.. Guilford Press, pp. 102–138. doi:10.1080/02699939208411068.

Joseph, R., Liu, T., Ng, A.B., See, S., Rai, S., 2023. NewsMet : A 'do it all' dataset of contemporary metaphors in news headlines, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada. pp. 10090–10104. URL: https://aclanthology.org/2023.findings-acl.641/, doi:10.18653/v1/2023.findings-acl.641.

Jung, C., Kim, D., Jin, J., Kim, J., Seonwoo, Y., Choi, Y., Oh, A., Kim, H., 2024. Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 19794–19809. URL: https://aclanthology.org/2024.emnlp-main.1105/, doi:10.18653/v1/2024.emnlp-main.1105.

Kang, M., Choi, G., Jeon, H., An, J.H., Choi, D., Han, J., 2024. CURE: Context- and uncertainty-aware mental disorder detection, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 17924–17940. URL: https://aclanthology.org/2024.emnlp-main.994/, doi:10.18653/v1/2024.emnlp-main.994.

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. Journal of neuroscience 17, 4302–4311.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D., 2020. Scaling laws for neural language models. URL: https://arxiv.org/abs/2001.08361, arXiv:2001.08361.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al., 2023. Chatgpt for good? on opportunities and challenges of large language models for education. Learning and individual differences 103, 102274.

Katz, D.M., Bommarito, M.J., Gao, S., Arredondo, P., 2024. Gpt-4 passes the bar exam. Philosophical Transactions of the Royal Society A 382, 20230254.

Kerlinger, F.N., 1966. Foundations of behavioral research. .

Kim, H., Sclar, M., Zhou, X., Bras, R., Kim, G., Choi, Y., Sap, M., 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore. pp. 14397–14413. URL: https://aclanthology.org/2023.emnlp-main.890/, doi:10.18653/v1/2023.emnlp-main.890.

Kim, Y., Park, C., Jeong, H., Chan, Y.S., Xu, X., McDuff, D., Lee, H., Ghassemi, M., Breazeal, C., Park, H.W., 2024. MDAgents: An adaptive collaboration of LLMs for medical decision-making, in: The Thirty-eighth Annual Conference on Neural Information Processing Systems. URL: https://openreview.net/forum?id=EKdk4vxK04.

Kirschner, P.A., Sweller, J., Clark, R.E., 2006. Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational psychologist 41, 75–86.

Kobayashi, M., Mita, M., Komachi, M., 2024. Revisiting meta-evaluation for grammatical error correction. Transactions of the Association for Computational Linguistics 12, 837–855. URL: https://aclanthology.org/2024.tacl-1.47/, doi:10.1162/tacl_a_00676.

Kong, Z., Goel, A., Badlani, R., Ping, W., Valle, R., Catanzaro, B., 2024. Audio flamingo: a novel audio language model with few-shot learning and dialogue abilities, in: Proceedings of the 41st International Conference on Machine Learning, JMLR.org.

Konrath, S., Meier, B.P., Bushman, B.J., 2018. Development and validation of the single item trait empathy scale (sites). Journal of research in personality 73, 111–122.

Koo, R., Lee, M., Raheja, V., Park, J.I., Kim, Z.M., Kang, D., 2024. Benchmarking cognitive biases in large language models as evaluators, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand. pp. 517–545. URL: https://aclanthology.org/2024.findings-acl.29/, doi:10.18653/v1/2024.findings-acl.29.

Kosinski, M., 2023. Theory of mind may have spontaneously emerged in large language models. arXiv preprint arXiv:2302.02083 4, 169.

Kraft, M.A., Blazar, D., Hogan, D., 2018. The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. Review of educational research 88, 547–588.

Krathwohl, D.R., 2002. A revision of bloom's taxonomy: An overview. Theory into practice 41, 212–218.

Kwon, D., Weiss, E., Kulshrestha, T., Chawla, K., Lucas, G., Gratch, J., 2024. Are LLMs effective negotiators? systematic evaluation of the multifaceted capabilities of LLMs in negotiation dialogues, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA. pp. 5391–5413. URL: https://aclanthology.org/2024.findings-emnlp.310/, doi:10.18653/v1/2024.findings-emnlp.310.

Lai, B., Zhang, H., Liu, M., Pariani, A., Ryan, F., Jia, W., Hayati, S.A., Rehg, J., Yang, D., 2023a. Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada. pp. 6570–6588. URL: https://aclanthology.org/2023.findings-acl.411/, doi:10.18653/v1/2023.findings-acl.411.

Lai, J., Gan, W., Wu, J., Qi, Z., Yu, P.S., 2023b. Large language models in law: A survey. URL: https://arxiv.org/abs/2312.03718, arXiv:2312.03718.

Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J., 2017. Building machines that learn and think like people. Behavioral and brain sciences 40, e253.

Laskar, M.T.R., Alqahtani, S., Bari, M.S., Rahman, M., Khan, M.A.M., Khan, H., Jahan, I., Bhuiyan, A., Tan, C.W., Parvez, M.R., Hoque, E., Joty, S., Huang, J., 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 13785–13816. URL: https://aclanthology.org/2024.emnlp-main.764/, doi:10.18653/v1/2024.emnlp-main.764.

Lazarus, R.S., 1966. Psychological stress and the coping process. .

LeDoux, J.E., 1998. The emotional brain: The mysterious underpinnings of emotional life. Simon and Schuster.

Lejuez, C.W., Read, J.P., Kahler, C.W., Richards, J.B., Ramsey, S.E., Stuart, G.L., Strong, D.R., Brown, R.A., 2002. Evaluation of a behavioral measure of risk taking: the balloon analogue risk task (bart). Journal of Experimental Psychology: Applied 8, 75.

Lester, B., Al-Rfou, R., Constant, N., 2021. The power of scale for parameter-efficient prompt tuning, in: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic. pp. 3045–3059. URL: https://aclanthology.org/2021.emnlp-main.243/, doi:10.18653/v1/2021.emnlp-main.243.

Levinson, S.C., 2000. Presumptive meanings: The theory of generalized conversational implicature. MIT press.

Levy, R., 2008. Expectation-based syntactic comprehension. Cognition 106, 1126–1177.

Li, C., Chen, M., Wang, J., Sitaram, S., Xie, X., 2024a. CultureLLM: Incorporating cultural differences into large language models, in: The Thirty-eighth Annual Conference on Neural Information Processing Systems. URL: https://openreview.net/forum?id=sIsbOkQmBL.

Li, C., Qi, Y., 2025. Toward accurate psychological simulations: Investigating llms' responses to personality and cultural variables. Computers in Human Behavior 170, 108687. URL: https://www.sciencedirect.com/science/article/pii/S0747563225001347, doi:https://doi.org/10.1016/j.chb.2025.108687.

Li, H., Chong, Y., Stepputtis, S., Campbell, J., Hughes, D., Lewis, C., Sycara, K., 2023. Theory of mind for multi-agent collaboration via large language models, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore. pp. 180–192. URL: https://aclanthology.org/2023.emnlp-main.13/, doi:10.18653/v1/2023.emnlp-main.13.

Li, J., Peris, C., Mehrabi, N., Goyal, P., Chang, K.W., Galstyan, A., Zemel, R., Gupta, R., 2024b. The steerability of large language models toward data-driven personas, in: Duh, K., Gomez, H., Bethard, S. (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico. pp. 7290–7305. URL: https://aclanthology.org/2024.naacl-long.405/, doi:10.18653/v1/2024.naacl-long.405.

Li, X.L., Liang, P., 2021. Prefix-tuning: Optimizing continuous prompts for generation, in: Zong, C., Xia, F., Li, W., Navigli, R. (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online. pp. 4582–4597. URL: https://aclanthology.org/2021.acl-long.353/, doi:10.18653/v1/2021.acl-long.353.

Li, Y., Chen, X., Hu, B., Shi, H., Zhang, M., 2024c. Cognitive visual-language mapper: Advancing multimodal comprehension with enhanced visual knowledge alignment, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 7615–7626. URL: https://aclanthology.org/2024.acl-long.411/, doi:10.18653/v1/2024.acl-long.411.

Li, Y., Ma, S., Wang, X., Huang, S., Jiang, C., Zheng, H.T., Xie, P., Huang, F., Jiang, Y., 2024d. Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 18582–18590.

Li, Z., Wang, C., Ma, P., Wu, D., Wang, S., Gao, C., Liu, Y., 2024e. Split and merge: Aligning position biases in LLM-based evaluators, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 11084–11108. URL: https://aclanthology.org/2024.emnlp-main.621/, doi:10.18653/v1/2024.emnlp-main.621.

Liao, K.Y.H., Wei, M., Yin, M., 2020. The misunderstood schema of the strong black woman: Exploring its mental health consequences and coping responses among african american women. Psychology of Women Quarterly 44, 84–104.

Likert, R., 1932. A technique for the measurement of attitudes. Archives of Psychology 22, 55.

Lin, Y., Lin, H., Xiong, W., Diao, S., Liu, J., Zhang, J., Pan, R., Wang, H., Hu, W., Zhang, H., Dong, H., Pi, R., Zhao, H., Jiang, N., Ji, H., Yao, Y., Zhang, T., 2024. Mitigating the alignment tax of RLHF, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 580–606. URL: https://aclanthology.org/2024.emnlp-main.35/, doi:10.18653/v1/2024.emnlp-main.35.

Lin, Z., Dai, Y., 2025. Fostering epistemic insights into ai ethics through a constructionist pedagogy: An interdisciplinary approach to ai literacy, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 29171–29177.

Lipton, Z.C., 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16, 31–57.

Lissak, S., Calderon, N., Shenkman, G., Ophir, Y., Fruchter, E., Brunstein Klomek, A., Reichart, R., 2024. The colorful future of LLMs: Evaluating and improving LLMs as emotional supporters for queer youth, in: Duh, K., Gomez, H., Bethard, S. (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico. pp. 2040–2079. URL: https://aclanthology.org/2024.naacl-long.113/, doi:10.18653/v1/2024.naacl-long.113.

Liu, A., Diab, M., Fried, D., 2024a. Evaluating large language model biases in persona-steered generation, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand. pp. 9832–9850. URL: https://aclanthology.org/2024.findings-acl.586/, doi:10.18653/v1/2024.findings-acl.586.

Liu, H., Li, C., Li, Y., Lee, Y.J., 2024b. Improved baselines with visual instruction tuning, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 26286–26296. doi:10.1109/CVPR52733.2024.02484.

Liu, J., Liu, H., Xiao, L., Wang, Z., Liu, K., Gao, S., Zhang, W., Zhang, S., Chen, K., 2025a. Are your llms capable of stable reasoning? URL: https://arxiv.org/abs/2412.13147, arXiv:2412.13147.

Liu, K., Fu, Z., Chen, C., Zhang, W., Jiang, R., Zhou, F., Chen, Y., Wu, Y., Ye, J., 2025b. Enhancing llm's cognition via structurization, in: Proceedings of the 38th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA.

Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P., 2024c. Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics 12, 157–173. URL: https://aclanthology.org/2024.tacl-1.9/, doi:10.1162/tacl_a_00638.

Liu, X., Su, K., Shlizerman, E., 2025c. Tell what you hear from what you see - video to audio generation through text, in: Proceedings of the 38th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA.

Liu, X., Yin, D., Zhang, C., Feng, Y., Zhao, D., 2023. The magic of IF: Investigating causal reasoning abilities in large language models of code, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada. pp. 9009–9022. URL: https://aclanthology.org/2023.findings-acl.574/, doi:10.18653/v1/2023.findings-acl.574.

Liu, Z., Gong, Z., Ai, L., Hui, Z., Chen, R., Leach, C.W., Greene, M.R., Hirschberg, J., 2025d. The mind in the machine: A survey of incorporating psychological theories in llms. arXiv preprint arXiv:2505.00003 .

Liu, Z., Oguz, B., Zhao, C., Chang, E., Stock, P., Mehdad, Y., Shi, Y., Krishnamoorthi, R., Chandra, V., 2024d. LLM-QAT: Data-free quantization aware training for large language models, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand. pp. 467–484. URL: https://aclanthology.org/2024.findings-acl.26/, doi:10.18653/v1/2024.findings-acl.26.

Lloyd, S., 1982. Least squares quantization in pcm. IEEE Transactions on Information Theory 28, 129–137. doi:10.1109/TIT.1982.1056489.

Lord, F., 1980. Applications of Item Response Theory To Practical Testing Problems. Routledge. URL: https://doi.org/10.4324/9780203056615.

Lord, F., Novick, M., Birnbaum, A., 1968. Statistical theories of mental test scores. Addison-Wesley, Oxford, England.

Luo, H., Deng, Y., Shen, Y., Ng, S.K., Chua, T.S., 2024. Chain-of-exemplar: Enhancing distractor generation for multimodal educational question generation, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 7978–7993. URL: https://aclanthology.org/2024.acl-long.432/, doi:10.18653/v1/2024.acl-long.432.

Ma, X., Fang, G., Wang, X., 2023. LLM-pruner: On the structural pruning of large language models, in: Thirty-seventh Conference on Neural Information Processing Systems. URL: https://openreview.net/forum?id=J8Ajf9WfXP.

Macina, J., Daheim, N., Chowdhury, S., Sinha, T., Kapur, M., Gurevych, I., Sachan, M., 2023. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 5602–5621. URL: https://aclanthology.org/2023.findings-emnlp.372/, doi:10.18653/v1/2023.findings-emnlp.372.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, University of California press. pp. 281–298.

Manakul, P., Liusie, A., Gales, M., 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore. pp. 9004–9017. URL: https://aclanthology.org/2023.emnlp-main.557/, doi:10.18653/v1/2023.emnlp-main.557.

Manvi, R., Khanna, S., Burke, M., Lobell, D., Ermon, S., 2024. Large language models are geographically biased, in: Proceedings of the 41st International Conference on Machine Learning, JMLR.org.

McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E., 2006. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. AI magazine 27, 12–12.

von Mengden, F., Coussé, E., 2014. Introduction. The role of change in usage-based conceptions of language. John Benjamins Publishing Company. pp. 1–20. URL: https://doi.org/10.1075/sfsl.69.01men, doi:doi:10.1075/sfsl.69.01men.

Meta, 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025-04-05.

Meyer, I.H., 2003. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. Psychological bulletin 129, 674.

Meyer, J.G., Urbanowicz, R.J., Martin, P.C., O'Connor, K., Li, R., Peng, P.C., Bright, T.J., Tatonetti, N., Won, K.J., Gonzalez-Hernandez, G., et al., 2023. Chatgpt and large language models in academia: opportunities and challenges. BioData mining 16, 20.

Microsoft, 2023. Microsoft 365 copilot release notes. https://learn.microsoft.com/en-us/copilot/microsoft-365/release-notes. Accessed: 2025-07-03.

Minixhofer, B., Pfeiffer, J., Vulić, I., 2023. Compoundpiece: Evaluating and improving decompounding performance of language models, in: The 2023 Conference on Empirical Methods in Natural Language Processing. URL: https://openreview.net/forum?id=xapBkUt0yf.

Miri, B., David, B.C., Uri, Z., 2007. Purposely teaching for the promotion of higher-order thinking skills: A case of critical thinking. Research in science education 37, 353–369.

Mishra, S., Lalumière, M.L., 2011. Individual differences in risk-propensity: Associations between personality and behavioral measures of risk. Personality and Individual Differences 50, 869–873.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C.D., Finn, C., 2023. Detectgpt: zero-shot machine-generated text detection using probability curvature, in: Proceedings of the 40th International Conference on Machine Learning, JMLR.org.

Moors, A., Ellsworth, P.C., Scherer, K.R., Frijda, N.H., 2013. Appraisal theories of emotion: State of the art and future development. Emotion review 5, 119–124.

Morabito, R., Madhusudan, S., McDonald, T., Emami, A., 2024. STOP! benchmarking large language models with sensitivity testing on offensive progressions, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 4221–4243. URL: https://aclanthology.org/2024.emnlp-main.243/, doi:10.18653/v1/2024.emnlp-main.243.

Munos, R., Valko, M., Calandriello, D., Azar, M.G., Rowland, M., Guo, Z.D., Tang, Y., Geist, M., Mesnard, T., Fiegel, C., Michi, A., Selvi, M., Girgin, S., Momchev, N., Bachem, O., Mankowitz, D.J., Precup, D., Piot, B., 2024. Nash learning from human feedback, in: Forty-first International Conference on Machine Learning. URL: https://openreview.net/forum?id=Y5AmNYiyCQ.

Nair, S., Resnik, P., 2023. Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship?, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 11251–11260. URL: https://aclanthology.org/2023.findings-emnlp.752/, doi:10.18653/v1/2023.findings-emnlp.752.

Nickerson, R.S., 1998. Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology 2, 175–220.

Nie, A., Zhang, Y., Amdekar, A., Piech, C.J., Hashimoto, T., Gerstenberg, T., 2023. Moca: Measuring human-language model alignment on causal and moral judgment tasks, in: Thirty-seventh Conference on Neural Information Processing Systems. URL: https://openreview.net/forum?id=UdByCgCNdr.

Niu, Z., Zhong, G., Yu, H., 2021. A review on the attention mechanism of deep learning. Neurocomputing 452, 48–62. URL: https://www.sciencedirect.com/science/article/pii/S092523122100477X, doi:https://doi.org/10.1016/j.neucom.2021.03.091.

Norman, D.A., Shallice, T., 1986. Attention to action: Willed and automatic control of behavior, in: Consciousness and self-regulation: Advances in research and theory volume 4. Springer, pp. 1–18.

Novick, M.R., 1966. The axioms and principal results of classical test theory. Journal of Mathematical Psychology 3, 1–18. URL: https://www.sciencedirect.com/science/article/pii/0022249666900022, doi:https://doi.org/10.1016/0022-2496(66)90002-2.

Oates, J.E., Grayson, A.E., 2004. Cognitive and language development in children. Open University Press.

Oliver, R.L., Balakrishnan, P.S., Barry, B., 1994. Outcome satisfaction in negotiation: A test of expectancy disconfirmation. Organizational behavior and human decision processes 60, 252–275.

Onishi, K.H., Baillargeon, R., 2005. Do 15-month-old infants understand false beliefs? science 308, 255–258.

OpenAI, 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-05-13.

OpenAI, 2025a. Gpt-4v(ision) system card. https://openai.com/index/gpt-4v-system-card/. Accessed: 2025-06-07.

OpenAI, 2025b. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed: 2025-04-16.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R., 2022. Training language models to follow instructions with human feedback. URL: https://arxiv.org/abs/2203.02155, arXiv:2203.02155.

Pan, J., Zhang, Y., Zhang, C., Liu, Z., Wang, H., Li, H., 2024. DynaThink: Fast or slow? a dynamic decision-making framework for large language models, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 14686–14695. URL: https://aclanthology.org/2024.emnlp-main.814/, doi:10.18653/v1/2024.emnlp-main.814.

Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.Y., Wang, W., 2023. On the risk of misinformation pollution with large language models, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 1389–1403. URL: https://aclanthology.org/2023.findings-emnlp.97/, doi:10.18653/v1/2023.findings-emnlp.97.

Persson, B.N., Kajonius, P.J., Garcia, D., 2019. Revisiting the structure of the short dark triad. Assessment 26, 3–16.

Piaget, J., Cook, M., et al., 1952. The origins of intelligence in children. volume 8. International universities press New York.

Pinker, S., 2003. The language instinct: How the mind creates language. Penguin uK.

Pintrich, P.R., 2002. The role of metacognitive knowledge in learning, teaching, and assessing. Theory into practice 41, 219–225.

Placani, A., 2024. Anthropomorphism in ai: Hype and fallacy. AI and Ethics 4, 691–698. URL: https://doi.org/10.1007/s43681-024-00419-4, doi:10.1007/s43681-024-00419-4.

Premack, D., Woodruff, G., 1978. Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences 1, 515–526. doi:10.1017/s0140525x00076512.

Qian, Y., Zhang, W., Liu, T., 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 6516–6528. URL: https://aclanthology.org/2023.findings-emnlp.433/, doi:10.18653/v1/2023.findings-emnlp.433.

Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Hong, L., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., Sun, M., 2024. Toolllm: Facilitating large language models to master 16000+ real-world apis, in: International Conference on Learning Representations.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C., 2023. Direct preference optimization: Your language model is secretly a reward model, in: Thirty-seventh Conference on Neural Information Processing Systems. URL: https://openreview.net/forum?id=HPuSIXJaa9.

Ratcliffe, M., 2006. 'folk psychology' is not folk psychology. Phenomenology and the Cognitive Sciences 5, 31–52. URL: https://doi.org/10.1007/s11097-005-9010-y, doi:10.1007/s11097-005-9010-y.

Reichardt, C.S., 2002. Experimental and quasi-experimental designs for generalized causal inference.

Ren, Y., Xiong, D., 2023. HuaSLIM: Human attention motivated shortcut learning identification and mitigation for large language models, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada. pp. 12350–12365. URL: https://aclanthology.org/2023.findings-acl.781/, doi:10.18653/v1/2023.findings-acl.781.

Roccas, S., Brewer, M.B., 2002. Social identity complexity. Personality and social psychology review 6, 88–106.

Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65, 386–408.

Rosenman, G., Hendler, T., Wolf, L., 2024. LLM questionnaire completion for automatic psychiatric assessment, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA. pp. 403–415. URL: https://aclanthology.org/2024.findings-emnlp.23/, doi:10.18653/v1/2024.findings-emnlp.23.

Rostam, Z.R.K., Szénási, S., Kertész, G., 2024. Achieving peak performance for large language models: A systematic review. IEEE Access 12, 96017–96050. doi:10.1109/ACCESS.2024.3424945.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence 1, 206–215.

Sabour, S., Liu, S., Zhang, Z., Liu, J., Zhou, J., Sunaryo, A., Lee, T., Mihalcea, R., Huang, M., 2024. EmoBench: Evaluating the emotional intelligence of large language models, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 5986–6004. URL: https://aclanthology.org/2024.acl-long.326/, doi:10.18653/v1/2024.acl-long.326.

Salemi, A., Mysore, S., Bendersky, M., Zamani, H., 2024. LaMP: When large language models meet personalization, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 7370–7392. URL: https://aclanthology.org/2024.acl-long.399/, doi:10.18653/v1/2024.acl-long.399.

Salter, L., Hearn, A., 1997. Outside the lines: Issues in interdisciplinary research. McGill-Queen's Press-MQUP.

Sanches, P., Janson, A., Karpashevich, P., Nadal, C., Qu, C., Daudén Roquet, C., Umair, M., Windlin, C., Doherty, G., Höök, K., Sas, C., 2019. Hci and affective health: Taking stock of a decade of studies and charting future research directions, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 1–17. URL: https://doi.org/10.1145/3290605.3300475, doi:10.1145/3290605.3300475.

Schacter, D.L., Gilbert, D.T., Wegner, D.M., 2009. Psychology. Macmillan.

Schaeffer, R., Miranda, B., Koyejo, S., 2023. Are emergent abilities of large language models a mirage? Advances in Neural Information Processing Systems 36, 55565–55581.

Scherer, K.R., Moors, A., 2019. The emotion process: Event appraisal and component differentiation. Annual review of psychology 70, 719–745.

Scherrer, N., Shi, C., Feder, A., Blei, D., 2023. Evaluating the moral beliefs encoded in LLMs, in: Thirty-seventh Conference on Neural Information Processing Systems. URL: https://openreview.net/forum?id=O06z2G18me.

Schultz, D., 2013. A history of modern psychology. Academic Press.

Schwartz, S.H., 2012. An overview of the schwartz theory of basic values. Online readings in Psychology and Culture 2, 11.

Shahaeian, A., Peterson, C.C., Slaughter, V., Wellman, H.M., 2011. Culture and the sequence of steps in theory of mind development. Developmental Psychology 47, 1239–1247. doi:https://doi.org/10.1037/a0023899.

Shaikh, O., Zhang, H., Held, W., Bernstein, M., Yang, D., 2023. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada. pp. 4454–4470. URL: https://aclanthology.org/2023.acl-long.244/, doi:10.18653/v1/2023.acl-long.244.

Shani, C., Vreeken, J., Shahaf, D., 2023. Towards concept-aware large language models, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 13158–13170. URL: https://aclanthology.org/2023.findings-emnlp.877/, doi:10.18653/v1/2023.findings-emnlp.877.

Shanton, K., Goldman, A., 2010. Simulation theory. Wiley Interdisciplinary Reviews: Cognitive Science 1, 527–538. doi:10.1002/wcs.33.

Shao, Y., Li, L., Dai, J., Qiu, X., 2023. Character-LLM: A trainable agent for role-playing, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore. pp. 13153–13187. URL: https://aclanthology.org/2023.emnlp-main.814/.

Shapira, N., Zwirn, G., Goldberg, Y., 2023. How well do large language models perform on faux pas tests?, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada. pp. 10438–10451. URL: https://aclanthology.org/2023.findings-acl.663/, doi:10.18653/v1/2023.findings-acl.663.

Shen, J., Mire, J., Park, H.W., Breazeal, C., Sap, M., 2024a. HEART-felt narratives: Tracing empathy and narrative style in personal stories with LLMs, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 1026–1046. URL: https://aclanthology.org/2024.emnlp-main.59/, doi:10.18653/v1/2024.emnlp-main.59.

Shen, S., Logeswaran, L., Lee, M., Lee, H., Poria, S., Mihalcea, R., 2024b. Understanding the capabilities and limitations of large language models for cultural commonsense, in: Duh, K., Gomez, H., Bethard, S. (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico. pp. 5668–5680. URL: https://aclanthology.org/2024.naacl-long.316/, doi:10.18653/v1/2024.naacl-long.316.

Shi, W., Li, R., Zhang, Y., Ziems, C., Yu, S., Horesh, R., Paula, R.A.D., Yang, D., 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA. pp. 4996–5025. URL: https://aclanthology.org/2024.findings-emnlp.288/, doi:10.18653/v1/2024.findings-emnlp.288.

Shiffrin, R.M., Nobel, P.A., 1997. The art of model development and testing. Behavior Research Methods, Instruments, & Computers 29, 6–14.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., y Arcas, B.A., Webster, D., Corrado, G.S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., Natarajan, V., 2023. Large language models encode clinical knowledge. Nature 620, 172–180. URL: https://doi.org/10.1038/s41586-023-06291-2, doi:10.1038/s41586-023-06291-2.

Skinner, B.F., 1965. Science and human behavior. 92904, Simon and Schuster.

Sonkar, S., Liu, N., Mallick, D., Baraniuk, R., 2023. CLASS: A design framework for building intelligent tutoring systems based on learning science principles, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 1941–1961. URL: https://aclanthology.org/2023.findings-emnlp.130/, doi:10.18653/v1/2023.findings-emnlp.130.

Sonkar, S., Ni, K., Chaudhary, S., Baraniuk, R., 2024. Pedagogical alignment of large language models, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA. pp. 13641–13650. URL: https://aclanthology.org/2024.findings-emnlp.797/, doi:10.18653/v1/2024.findings-emnlp.797.

Soydaner, D., 2022. Attention mechanism in neural networks: where it comes and where it goes. Neural Computing and Applications 34, 13371–13385. URL: https://doi.org/10.1007/s00521-022-07366-3, doi:10.1007/s00521-022-07366-3.

Spearman, C., 2010. The proof and measurement of association between two things. International Journal of Epidemiology 39, 1137–1150. URL: https://doi.org/10.1093/ije/dyq191, doi:10.1093/ije/dyq191, arXiv:https://academic.oup.com/ije/article-pdf/39/5/1137/18481215/dyq191.pdf.

Spitzer, R.L., Kroenke, K., Williams, J.B., Löwe, B., 2006. A brief measure for assessing generalized anxiety disorder: the gad-7. Archives of internal medicine 166, 1092–1097.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al., 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research URL: https://openreview.net/forum?id=uyTL5Bvosj. featured Certification.

Stipek, D., Iver, D.M., 1989. Developmental change in children's assessment of intellectual competence. Child development , 521–538.

Stivers, T., Enfield, N.J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J.P., Yoon, K.E., Levinson, S.C., 2009. Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences 106, 10587–10592. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0903616106, doi:10.1073/pnas.0903616106, arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.0903616106.

Street, W., 2024. Llm theory of mind and alignment: Opportunities and risks, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems.

Sui, P., Duede, E., Wu, S., So, R., 2024. Confabulation: The surprising value of large language model hallucinations, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 14274–14284. URL: https://aclanthology.org/2024.acl-long.770/, doi:10.18653/v1/2024.acl-long.770.

Sullivan, M.E., Yates, K.A., Inaba, K., Lam, L., Clark, R.E., 2014. The use of cognitive task analysis to reveal the instructional limitations of experts in the teaching of procedural skills. Academic Medicine 89, 811–816.

Sun, S., Lee, E., Baek, S.Y., Hwang, S., Lee, W., Nan, D., Jansen, B.J., Kim, J.H., 2024. Kiss up, kick down: Exploring behavioral changes in multi-modal large language models with assigned visual personas, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 10888–10901. URL: https://aclanthology.org/2024.emnlp-main.609/, doi:10.18653/v1/2024.emnlp-main.609.

Suresh, S., Mukherjee, K., Yu, X., Huang, W.C., Padua, L., Rogers, T., 2023. Conceptual structure coheres in human cognition but not in large language models, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore. pp. 722–738. URL: https://aclanthology.org/2023.emnlp-main.47/, doi:10.18653/v1/2023.emnlp-main.47.

Tan, Z., Liu, Z., Jiang, M., 2024. Personalized pieces: Efficient personalized large language models through collaborative efforts, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 6459–6475. URL: https://aclanthology.org/2024.emnlp-main.371/, doi:10.18653/v1/2024.emnlp-main.371.

Tang, T., Luo, W., Huang, H., Zhang, D., Wang, X., Zhao, X., Wei, F., Wen, J.R., 2024. Language-specific neurons: The key to multilingual capabilities in large language models, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 5701–5715. URL: https://aclanthology.org/2024.acl-long.309/, doi:10.18653/v1/2024.acl-long.309.

Tenenbaum, J.B., Griffiths, T.L., Kemp, C., 2006. Theory-based bayesian models of inductive learning and reasoning. Trends in cognitive sciences 10, 309–318.

Tennenholtz, G., Chow, Y., Hsu, C., Jeong, J., Shani, L., Tulepbergenov, A., Ramachandran, D., Mladenov, M., Boutilier, C., 2024. Demystifying embedding spaces using large language models, in: The Twelfth International Conference on Learning Representations. URL: https://openreview.net/forum?id=qoYogklIPz.

Testa, D., Chersoni, E., Lenci, A., 2023. We understand elliptical sentences, and language models should too: A new dataset for studying ellipsis and its interaction with thematic fit, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada. pp. 3340–3353. URL: https://aclanthology.org/2023.acl-long.188/, doi:10.18653/v1/2023.acl-long.188.

Tian, Y., Xu, N., Mao, W., 2024. A theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning, in: Duh, K., Gomez, H., Bethard, S. (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico. pp. 7738–7755. URL: https://aclanthology.org/2024.naacl-long.428/, doi:10.18653/v1/2024.naacl-long.428.

Towner, S., 2010. Concept of mind in non-human primates. Bioscience Horizons: The International Journal of Student Research 3, 96–104. URL: https://doi.org/10.1093/biohorizons/hzq011, doi:10.1093/biohorizons/hzq011, arXiv:https://academic.oup.com/biohorizons/article-pdf/3/1/96/5031707/hzq011.pdf.

Treisman, A.M., 1964. Selective attention in man. British medical bulletin .

Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: Heuristics and biases. Science 185, 1124–1131. URL: https://www.science.org/doi/abs/10.1126/science.185.4157.1124, doi:10.1126/science.185.4157.1124, arXiv:https://www.science.org/doi/pdf/10.1126/science.185.4157.1124.

Tversky, A., Kahneman, D., 1981. The framing of decisions and the psychology of choice. Science 211, 453–458. URL: https://www.science.org/doi/abs/10.1126/science.7455683, doi:10.1126/science.7455683, arXiv:https://www.science.org/doi/pdf/10.1126/science.7455683.

Valmeekam, K., Marquez, M., Sreedharan, S., Kambhampati, S., 2023. On the planning abilities of large language models - a critical investigation, in: Thirty-seventh Conference on Neural Information Processing Systems. URL: https://openreview.net/forum?id=X6dEqXIsEW.

de Varda, A., Marelli, M., 2023. Scaling in cognitive modelling: a multilingual approach to human reading times, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Toronto, Canada. pp. 139–149. URL: https://aclanthology.org/2023.acl-short.14/, doi:10.18653/v1/2023.acl-short.14.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Wachowiak, L., Gromann, D., 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada. pp. 1018–1032. URL: https://aclanthology.org/2023.acl-long.58/, doi:10.18653/v1/2023.acl-long.58.

Walker, S.P., 2016. Thinking, straight or true? https://steer.education/scientific-research/thinking-straight-or-true/. Accessed: 2025-07-23.

Wang, A., Yin, Z., Hu, Y., Mao, Y., Hui, P., 2024a. Exploring the potential of large language models in artistic creation: Collaboration and reflection on creative programming. URL: https://arxiv.org/abs/2402.09750, arXiv:2402.09750.

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., Truong, S.T., Arora, S., Mazeika, M., Hendrycks, D., Lin, Z., Cheng, Y., Koyejo, S., Song, D., Li, B., 2023a. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models, in: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track. URL: https://openreview.net/forum?id=kaHpo8OZw2.

Wang, J.X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J.Z., Hassabis, D., Botvinick, M., 2018. Prefrontal cortex as a meta-reinforcement learning system. Nature neuroscience 21, 860–868.

Wang, M., Yao, Y., Xu, Z., Qiao, S., Deng, S., Wang, P., Chen, X., Gu, J.C., Jiang, Y., Xie, P., Huang, F., Chen, H., Zhang, N., 2024b. Knowledge mechanisms in large language models: A survey and perspective, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA. pp. 7097–7135. URL: https://aclanthology.org/2024.findings-emnlp.416/, doi:10.18653/v1/2024.findings-emnlp.416.

Wang, R., Milani, S., Chiu, J.C., Zhi, J., Eack, S.M., Labrum, T., Murphy, S.M., Jones, N., Hardy, K.V., Shen, H., Fang, F., Chen, Z., 2024c. PATIENT-ψ: Using large language models to simulate patients for training mental health professionals, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 12772–12797. URL: https://aclanthology.org/2024.emnlp-main.711/, doi:10.18653/v1/2024.emnlp-main.711.

Wang, R., Zelikman, E., Poesia, G., Pu, Y., Haber, N., Goodman, N., 2024d. Hypothesis search: Inductive reasoning with language models, in: The Twelfth International Conference on Learning Representations. URL: https://openreview.net/forum?id=G7UtIGQmjm.

Wang, R., Zhang, Q., Robinson, C., Loeb, S., Demszky, D., 2024e. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes, in: Duh, K., Gomez, H., Bethard, S. (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico. pp. 2174–2199. URL: https://aclanthology.org/2024.naacl-long.120/, doi:10.18653/v1/2024.naacl-long.120.

Wang, W., Dong, L., Cheng, H., Liu, X., Yan, X., Gao, J., Wei, F., 2023b. Augmenting language models with long-term memory, in: Thirty-seventh Conference on Neural Information Processing Systems. URL: https://openreview.net/forum?id=BryMFPQ4L6.

Wang, Y., Duan, J., Fox, D., Srinivasa, S., 2023c. NEWTON: Are large language models capable of physical reasoning?, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 9743–9758. URL: https://aclanthology.org/2023.findings-emnlp.652/, doi:10.18653/v1/2023.findings-emnlp.652.

Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X., Liu, Q., 2023d. Aligning large language models with human: A survey. URL: https://arxiv.org/abs/2307.12966, arXiv:2307.12966.

Wang, Z., Bi, B., Pentyala, S.K., Ramnath, K., Chaudhuri, S., Mehrotra, S., Zixu, Zhu, Mao, X.B., Asur, S., Na, Cheng, 2024f. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. URL: https://arxiv.org/abs/2407.16216, arXiv:2407.16216.

Wang, Z., Dong, Y., Zeng, J., Adams, V., Sreedhar, M.N., Egert, D., Delalleau, O., Scowcroft, J., Kant, N., Swope, A., Kuchaiev, O., 2024g. HelpSteer: Multi-attribute helpfulness dataset for SteerLM, in: Duh, K., Gomez, H., Bethard, S. (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico. pp. 3371–3384. URL: https://aclanthology.org/2024.naacl-long.185/, doi:10.18653/v1/2024.naacl-long.185.

Wang, Z., Mao, S., Wu, W., Ge, T., Wei, F., Ji, H., 2024h. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration, in: Duh, K., Gomez, H., Bethard, S. (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico. pp. 257–279. URL: https://aclanthology.org/2024.naacl-long.15/, doi:10.18653/v1/2024.naacl-long.15.

Watson, J.B., 1913. Psychology as the behaviorist views it. Psychological review 20, 158.

Wei, A., Haghtalab, N., Steinhardt, J., 2023a. Jailbroken: How does LLM safety training fail?, in: Thirty-seventh Conference on Neural Information Processing Systems. URL: https://openreview.net/forum?id=jA235JGM09.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V., 2022a. Finetuned language models are zero-shot learners, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=gEZrGCozdqR.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al., 2022b. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 .

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D., 2023b. Chain-of-thought prompting elicits reasoning in large language models. URL: https://arxiv.org/abs/2201.11903, arXiv:2201.11903.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L.A., Isaac, W., Legassick, S., Irving, G., Gabriel, I., 2021. Ethical and social risks of harm from language models. arXiv:2112.04359.

Wellman, H.M., Liu, D., 2004. Scaling of theory-of-mind tasks. Child development 75, 523–541.

Wicke, P., Wachowiak, L., 2024. Exploring spatial schema intuitions in large language and vision models, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand. pp. 6102–6117. URL: https://aclanthology.org/2024.findings-acl.365/, doi:10.18653/v1/2024.findings-acl.365.

Wijesiriwardene, T., Wickramarachchi, R., Gajera, B., Gowaikar, S., Gupta, C., Chadha, A., Reganti, A.N., Sheth, A., Das, A., 2023. ANALOGICAL - a novel benchmark for long text analogy evaluation in large language models, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada. pp. 3534–3549. URL: https://aclanthology.org/2023.findings-acl.218/, doi:10.18653/v1/2023.findings-acl.218.

Wilf, A., Lee, S., Liang, P.P., Morency, L.P., 2024. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 8292–8308. URL: https://aclanthology.org/2024.acl-long.451/, doi:10.18653/v1/2024.acl-long.451.

Wimmer, H., Perner, J., 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 13, 103–128. URL: https://www.sciencedirect.com/science/article/pii/0010027783900045, doi:https://doi.org/10.1016/0010-0277(83)90004-5.

Wolf, Y., Wies, N., Avnery, O., Levine, Y., Shashua, A., 2024. Fundamental limitations of alignment in large language models, in: Proceedings of the 41st International Conference on Machine Learning, JMLR.org.

Wu, D., Shi, H., Sun, Z., Liu, B., 2024a. Deciphering digital detectives: Understanding LLM behaviors and capabilities in multi-agent mystery games, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand. pp. 8225–8291. URL: https://aclanthology.org/2024.findings-acl.490/, doi:10.18653/v1/2024.findings-acl.490.

Wu, W., Mao, S., Zhang, Y., Xia, Y., Dong, L., Cui, L., Wei, F., 2024b. Mind's eye of LLMs: Visualization-of-thought elicits spatial reasoning in large language models, in: The Thirty-eighth Annual Conference on Neural Information Processing Systems. URL: https://openreview.net/forum?id=CEJ1mYPgWw.

Wu, Y., He, Y., Jia, Y., Mihalcea, R., Chen, Y., Deng, N., 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 10691–10706. URL: https://aclanthology.org/2023.findings-emnlp.717/, doi:10.18653/v1/2023.findings-emnlp.717.

Wundt, W.M., 1904. Principles of physiological psychology. volume 1. Sonnenschein.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huang, X., Gui, T., 2023. The rise and potential of large language model based agents: A survey. arXiv:2309.07864.

Xiao, M., Xie, Q., Kuang, Z., Liu, Z., Yang, K., Peng, M., Han, W., Huang, J., 2024. HealMe: Harnessing cognitive reframing in large language models for psychotherapy, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 1707–1725. URL: https://aclanthology.org/2024.acl-long.93/, doi:10.18653/v1/2024.acl-long.93.

Xiao, Z., Zhang, S., Lai, V., Liao, Q.V., 2023. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore. pp. 10967–10982. URL: https://aclanthology.org/2023.emnlp-main.676/, doi:10.18653/v1/2023.emnlp-main.676.

Xie, J., Zhang, K., Chen, J., Lou, R., Su, Y., 2024a. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts, in: The Twelfth International Conference on Learning Representations. URL: https://openreview.net/forum?id=auKAUJZMO6.

Xie, J., Zhang, K., Chen, J., Zhu, T., Lou, R., Tian, Y., Xiao, Y., Su, Y., 2024b. Travelplanner: a benchmark for real-world planning with language agents, in: Proceedings of the 41st International Conference on Machine Learning, JMLR.org.

Xu, H., Zhao, R., Zhu, L., Du, J., He, Y., 2024a. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 8593–8623. URL: https://aclanthology.org/2024.acl-long.466/, doi:10.18653/v1/2024.acl-long.466.

Xu, J., Fei, H., Pan, L., Liu, Q., Lee, M.L., Hsu, W., 2024b. Faithful logical reasoning via symbolic chain-of-thought, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 13326–13365. URL: https://aclanthology.org/2024.acl-long.720/, doi:10.18653/v1/2024.acl-long.720.

Xu, L., Hu, Z., Zhou, D., Ren, H., Dong, Z., Keutzer, K., Ng, S.K., Feng, J., 2024c. MAgIC: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 7315–7332. URL: https://aclanthology.org/2024.emnlp-main.416/, doi:10.18653/v1/2024.emnlp-main.416.

Xu, R., Lin, B., Yang, S., Zhang, T., Shi, W., Zhang, T., Fang, Z., Xu, W., Qiu, H., 2024d. The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 16259–16303. URL: https://aclanthology.org/2024.acl-long.858/, doi:10.18653/v1/2024.acl-long.858.

Xu, R., Qi, Z., Guo, Z., Wang, C., Wang, H., Zhang, Y., Xu, W., 2024e. Knowledge conflicts for LLMs: A survey, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 8541–8565. URL: https://aclanthology.org/2024.emnlp-main.486/, doi:10.18653/v1/2024.emnlp-main.486.

Xu, R., Zhou, Z., Zhang, T., Qi, Z., Yao, S., Xu, K., Xu, W., Qiu, H., 2024f. Walking in others' shoes: How perspective-taking guides large language models in reducing toxicity and bias, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 8341–8368. URL: https://aclanthology.org/2024.emnlp-main.476/, doi:10.18653/v1/2024.emnlp-main.476.

Yang, H., Zhang, Y., Xu, J., Lu, H., Heng, P.A., Lam, W., 2024. Unveiling the generalization power of fine-tuned large language models, in: Duh, K., Gomez, H., Bethard, S. (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico. pp. 884–899. URL: https://aclanthology.org/2024.naacl-long.51/, doi:10.18653/v1/2024.naacl-long.51.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., Narasimhan, K.R., 2023. Tree of thoughts: Deliberate problem solving with large language models, in: Thirty-seventh Conference on Neural Information Processing Systems. URL: https://openreview.net/forum?id=5Xc1ecxO1h.

Yen, H., Gao, T., Chen, D., 2024. Long-context language modeling with parallel context encoding, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 2588–2610. URL: https://aclanthology.org/2024.acl-long.142/, doi:10.18653/v1/2024.acl-long.142.

Yi, J., Ye, R., Chen, Q., Zhu, B., Chen, S., Lian, D., Sun, G., Xie, X., Wu, F., 2024. On the vulnerability of safety alignment in open-access LLMs, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand. pp. 9236–9260. URL: https://aclanthology.org/2024.findings-acl.549/, doi:10.18653/v1/2024.findings-acl.549.

Yin, F., Vig, J., Laban, P., Joty, S., Xiong, C., Wu, C.S., 2023. Did you read the instructions? rethinking the effectiveness of task definitions in instruction learning, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada. pp. 3063–3079. URL: https://aclanthology.org/2023.acl-long.172/, doi:10.18653/v1/2023.acl-long.172.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E., 2024. A survey on multimodal large language models. National Science Review 11, nwae403. URL: https://doi.org/10.1093/nsr/nwae403, doi:10.1093/nsr/nwae403, arXiv:https://academic.oup.com/nsr/article-pdf/11/12/nwae403/61201557/nwae403.pdf.

Ying, J., Cao, Y., Bai, Y., Sun, Q., Wang, B., Tang, W., Ding, Z., Yang, Y., Huang, X., YAN, S., 2024a. Automating dataset updates towards reliable and timely evaluation of large language models, in: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track. URL: https://openreview.net/forum?id=EvEqYlQv8T.

Ying, J., Lin, M., Cao, Y., Tang, W., Wang, B., Sun, Q., Huang, X., Yan, S., 2024b. LLMs-as-instructors: Learning from errors toward automating model improvement, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA. pp. 11185–11208. URL: https://aclanthology.org/2024.findings-emnlp.654/, doi:10.18653/v1/2024.findings-emnlp.654.

Yona, G., Aharoni, R., Geva, M., 2024. Can large language models faithfully express their intrinsic uncertainty in words?, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 7752–7764. URL: https://aclanthology.org/2024.emnlp-main.443/, doi:10.18653/v1/2024.emnlp-main.443.

Yu, J., Wang, X., Tu, S., Cao, S., Zhang-Li, D., Lv, X., Peng, H., Yao, Z., Zhang, X., Li, H., Li, C., Zhang, Z., Bai, Y., Liu, Y., Xin, A., Yun, K., GONG, L., Lin, N., Chen, J., Wu, Z., Qi, Y., Li, W., Guan, Y., Zeng, K., Qi, J., Jin, H., Liu, J., Gu, Y., Yao, Y., Ding, N., Hou, L., Liu, Z., Bin, X., Tang, J., Li, J., 2024a. KoLA: Carefully benchmarking world knowledge of large language models, in: The Twelfth International Conference on Learning Representations. URL: https://openreview.net/forum?id=AqN23oqraW.

Yu, L., Yu, B., Yu, H., Huang, F., Li, Y., 2024b. Language models are super mario: absorbing abilities from homologous models as a free lunch, in: Proceedings of the 41st International Conference on Machine Learning, JMLR.org.

Yuan, R., Lin, H., Wang, Y., Tian, Z., Wu, S., Shen, T., Zhang, G., Wu, Y., Liu, C., Zhou, Z., Xue, L., Ma, Z., Liu, Q., Zheng, T., Li, Y., Ma, Y., Liang, Y., Chi, X., Liu, R., Wang, Z., Lin, C., Liu, Q., Jiang, T., Huang, W., Chen, W., Fu, J., Benetos, E., Xia, G., Dannenberg, R., Xue, W., Kang, S., Guo, Y., 2024. ChatMusician: Understanding and generating music intrinsically with LLM, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand. pp. 6252–6271. URL: https://aclanthology.org/2024.findings-acl.373/, doi:10.18653/v1/2024.findings-acl.373.

Zeng, Y., Lin, H., Zhang, J., Yang, D., Jia, R., Shi, W., 2024. How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 14322–14350. URL: https://aclanthology.org/2024.acl-long.773/, doi:10.18653/v1/2024.acl-long.773.

Zhang, C., Jian, Y., Ouyang, Z., Vosoughi, S., 2024a. Working memory identifies reasoning limits in language models, in: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 16896–16922. URL: https://aclanthology.org/2024.emnlp-main.938/, doi:10.18653/v1/2024.emnlp-main.938.

Zhang, J., Xu, X., Zhang, N., Liu, R., Hooi, B., Deng, S., 2024b. Exploring collaboration mechanisms for LLM agents: A social psychology view, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 14544–14607. URL: https://aclanthology.org/2024.acl-long.782/, doi:10.18653/v1/2024.acl-long.782.

Zhang, R., Cahyawijaya, S., Cruz, J.C.B., Winata, G., Aji, A.F., 2023a. Multilingual large language models are not (yet) code-switchers, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore. pp. 12567–12582. URL: https://aclanthology.org/2023.emnlp-main.774/, doi:10.18653/v1/2023.emnlp-main.774.

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., Wang, G., 2024c. Instruction tuning for large language models: A survey. URL: https://arxiv.org/abs/2308.10792, arXiv:2308.10792.

Zhang, X., Li, S., Hauer, B., Shi, N., Kondrak, G., 2023b. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore. pp. 7915–7927. URL: https://aclanthology.org/2023.emnlp-main.491/, doi:10.18653/v1/2023.emnlp-main.491.

Zhang, Z., Sheng, Y., Zhou, T., Chen, T., Zheng, L., Cai, R., Song, Z., Tian, Y., Re, C., Barrett, C., Wang, Z., Chen, B., 2023c. H2o: Heavy-hitter oracle for efficient generative inference of large language models, in: Thirty-seventh Conference on Neural Information Processing Systems. URL: https://openreview.net/forum?id=RkRrPp7GKO.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., Du, M., 2024a. Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology 15, 1–38.

Zhao, R., Zhu, Q., Xu, H., Li, J., Zhou, Y., He, Y., Gui, L., 2024b. Large language models fall short: Understanding complex relationships in detective narratives, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand. pp. 7618–7638. URL: https://aclanthology.org/2024.findings-acl.454/, doi:10.18653/v1/2024.findings-acl.454.

Zhen, H., Qiu, X., Chen, P., Yang, J., Yan, X., Du, Y., Hong, Y., Gan, C., 2024. 3d-vla: a 3d vision-language-action generative world model, in: Proceedings of the 41st International Conference on Machine Learning, JMLR.org.

Zheng, H.S., Mishra, S., Chen, X., Cheng, H.T., Chi, E.H., Le, Q.V., Zhou, D., 2024. Take a step back: Evoking reasoning via abstraction in large language models, in: The Twelfth International Conference on Learning Representations. URL: https://openreview.net/forum?id=3bq3jsvcQ1.

Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I., 2023. Judging llm-as-a-judge with mt-bench and chatbot arena, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA.

Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., Duan, N., 2024. AGIEval: A human-centric benchmark for evaluating foundation models, in: Duh, K., Gomez, H., Bethard, S. (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico. pp. 2299–2314. URL: https://aclanthology.org/2024.findings-naacl.149/, doi:10.18653/v1/2024.findings-naacl.149.

Zhou, H., Qian, J., Feng, Z., Hui, L., Zhu, Z., Mao, K., 2024a. LLMs learn task heuristics from demonstrations: A heuristic-driven prompting strategy for document-level event argument extraction, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 11972–11990. URL: https://aclanthology.org/2024.acl-long.647/, doi:10.18653/v1/2024.acl-long.647.

Zhou, X., Zhu, H., Mathur, L., Zhang, R., Yu, H., Qi, Z., Morency, L.P., Bisk, Y., Fried, D., Neubig, G., Sap, M., 2024b. SOTOPIA: Interactive evaluation for social intelligence in language agents, in: The Twelfth International Conference on Learning Representations. URL: https://openreview.net/forum?id=mM7VurbA4r.

Zhu, K., Chen, J., Wang, J., Gong, N.Z., Yang, D., Xie, X., 2024a. Dyval: Graph-informed dynamic evaluation of large language models, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=gjfOL9z5Xr.

Zhu, W., Zhang, Z., Wang, Y., 2024b. Language models represent beliefs of self and others, in: Proceedings of the 41st International Conference on Machine Learning, JMLR.org.

Zimmerman, C., 2000. The development of scientific reasoning skills. Developmental review 20, 99–149.

Cluster Summarization Template

```
You are an expert in literature review.                                    System prompt
-------------------------------------------------------------------------------------
Summarize five key phrases from the following paper titles and abstracts:  User prompt

# Title
{Title of paper in the cluster}
# Abstract
{Abstract of paper in the cluster} *20

The five key phrases for these papers should be (OUTPUT ONLY PHRASES SEPARATED BYCOMMAS):
```

Figure 12: Instruction Template for GPT-4o to Summarize Key Phrases in a Paper Cluster.

Theory/Framework Extraction Template

```
You are an expert in psychology literature review.                          System prompt
-------------------------------------------------------------------------------------
Identify the five most prominent psychological theories and frameworks related  User prompt
to [{primary psychology cluster name}] present in the following paper titles and
abstracts:

# Title
{Title of paper in a secondary psychology cluster}
# Abstract
{Abstract of paper in a secondary psychology cluster} *20

Merge identical theories or frameworks that appear under different names; do not provide
explanations.
```

Figure 13: Instruction Template for GPT-4.1 to Extract Candidate Psychology Theories and Frameworks from a Secondary Psychology Cluster.

A Instructions for GPT

In the citation analysis, we used GPT-4o to derive cluster topics and GPT-4.1 to extract and connect psychology theories and frameworks, as mentioned in §3.2 and §3.3, respectively. The instruction templates are provided in Fig. 12, Fig. 13, and Fig. 14.

Theory/Framework Connection Template

System prompt

```
You are an expert in psychology literature review.
```

User prompt

```
Here is information of a psychology paper:

# Title
{Title of the paper in a primary psychology cluster}
# Abstract
{Abstract of the paper in a primary psychology cluster}

Does this paper involve any of the following psychological theories or frameworks?
1. {Theory/framework in a secondary psychology cluster}
2. {Theory/framework in a secondary psychology cluster}
3. {Theory/framework in a secondary psychology cluster}

Answer in the following JSON format:
```
{"1": "[Y/N]", "2": "[Y/N]", "3": "[Y/N]"}
```
```

Figure 14: Instruction Template for GPT-4.1 to Link a Psychology Paper with Theories and Frameworks in a Secondary Psychology Cluster.