# A Review of Incorporating Psychological Theories in LLMs

**Anonymous ACL submission**


## Abstract

Psychological insights have long shaped pivotal NLP breakthroughs, from attention mechanisms to reinforcement learning and social modeling. As Large Language Models (LLMs) develop, there is a rising consensus that psychology is essential for capturing human-like cognition, behavior, and interaction. This paper reviews how psychological theories can inform and enhance stages of LLM development. Our review integrates insights from six subfields of psychology, highlighting current trends and gaps in how psychological theories are applied. By examining both cross-domain connections and points of tension, we aim to bridge disciplinary divides and promote more thoughtful integration of psychology into NLP research.


## 1 Introduction

As Large Language Models (LLMs) grow in scale and complexity, the Natural Language Processing (NLP) community increasingly sees psychology as key to capturing human-like cognition, behavior, and interaction (Qu et al., 2024; Lewis, 2025). Psychology, grounded in empirically validated and computationally adaptable frameworks (Sartori and Orrù, 2023; Ong, 2024), can address core LLM challenges such as reasoning fidelity, context retention, and user interaction. Reflecting these strengths, psychological insights have driven NLP advances, including the cognitive inspirations of attention mechanisms, formative reinforcement learning approaches, and social modeling for agents.

Despite extensive multidisciplinary efforts, a holistic review systematically integrating psychology across the LLM lifecycle remains missing. Most surveys and position papers remain fragmented, typically falling into three broad categories: **(1)** Some investigate how LLMs can empower traditional psychology or cognitive science research, for instance by modeling human reasoning and behavior at scale (Demszky et al., 2023; Abdurahman et al., 2024; Ong, 2024; Ke et al., 2024). **(2)** Others approach LLMs as subjects of psychological analysis, aiming to adapt or extend psychological theory, such as personality or cognition frameworks, to interpret and evaluate model behavior (Li et al., 2024b; Hagendorff et al., 2023; tse Huang et al., 2024; Pellert et al., 2024). **(3)** Finally, a third group leverages a single or limited set of psychological constructs to enhance model alignment or multi-agent frameworks – improving system reliability, social interaction, and trustworthiness (Liu et al., 2023; Dong et al., 2024b). This includes research on social influence for AI safety (Zeng et al., 2024a), moral reasoning in legal tasks (Almeida et al., 2024), and partial integrations of social or developmental psychology (Sartori and Orrù, 2023; Zhang et al., 2024c; Serapio-García et al., 2025). However, no existing work provides a unified map of how diverse psychological sub-areas can be harnessed, from data through application. Our survey fills this gap by offering a stage-wise view of how psychology can strengthen LLM capabilities and alignment across the entire lifecycle.

To address this gap, we present a structured review that situates psychological theories from six major areas across the entire LLM development pipeline. The contribution of our survey[1] are twofold: **(1)** We systematically review psychological theories applied in key stages of LLM development, identifying gaps and inconsistencies. **(2)** We highlight under-explored concepts alongside critical issues and debates at the intersection of psychology and NLP. Collectively, these contributions demonstrate how integrating diverse psychological frameworks can strengthen LLM design, enhance alignment, and broaden the practical and ethical impact of modern NLP systems.

As shown in Figure 1, the remainder of this pa-

[1] We survey 227 papers from major *CL venues, plus COLING, NeurIPS, ICML, ICLR, and influential arXiv preprints. Appendix B details paper selection process.

per illustrates how cognitive, developmental, behavioral, social, psycholinguistic, and personality theories integrate into four key stages of LLM development: preprocessing (Section 2), pre-training (Section 3), post-training (Section 4), and evaluation and application (Section 5). Finally, Section 6 discusses three central questions: *How does current LLM development leverage psychological theories? Which untapped psychological insights could advance LLM development? And what debates loom at the intersection of NLP and psychology?*

## 2 Preprocessing

We begin the stage-by-stage analysis of LLM development with preprocessing, the foundation that shapes downstream capabilities. Psychology provides valuable frameworks for understanding how humans acquire and filter information, underscoring the need for realistic, developmentally informed datasets and effective filtering strategies.

**Data Construction** Recent evidence shows that LLMs can align with human brain responses under biologically plausible training conditions (Hosseini et al., 2024), despite LLMs typically requiring orders of magnitude more training data than human. This supports the application of ***ecological validity*** (Schmuckler, 2001) that ***emphasizes real-world data to mimic cognitive development***. To reflect children's language acquisition processes, Jagadish et al. (2024) selects linguistically diverse environments, Feng et al. (2024) utilizes child-directed speech, while Nikolaus et al. (2022b) collects child cartoon. In parallel, ***incremental numerical understanding*** (Piaget, 2013) that ***views numerical concepts as gradually acquired through exposure*** is applied to sequential data collection with mathematically coherent numeric anchors (Sharma et al., 2024). Lastly, Reuben et al. (2025) provides a systematic framework to reformulate psychological questionnaires for LLM assessment.

**Data Preprocessing** Data preprocessing inspired by cognitive psychology involves refining data to enhance informational coherence prior to training. ***Selective attention*** (Treisman, 1969), ***prioritizing salient information while filtering out irrelevant stimuli***, was implemented to develop a preprocessing model that filters irrelevant data (Nottingham et al., 2024). Meanwhile, ***predictive coding*** proposing ***anticipatory processing based on prior knowledge*** (Rao and Ballard, 1999), was leveraged by Araujo et al. (2021) to enable anticipation of subsequent content, improving semantic coherence through expectation-driven processing. Lastly, drawing insights from ***knowledge acquisition of children***, Ficarra et al. (2025) redefines lexical knowledge in data to capture distributional information based on target word.

## 3 Pre-Training

Building on the foundations established during preprocessing, pre-training mirrors human cognitive development, where linguistic and reasoning abilities emerge through exposure to stimuli. This section explores how psychology inform observational learning and knowledge acquisition in LLMs.

**Observational Learning** ***Incremental cognitive development*** (Piaget, 1976), which posits ***children acquire knowledge through sequential tasks***, informs how LLMs can master nuanced concepts with explicit structured exposure. This principle manifests in Schulze Buschoff et al. (2023)'s gradually expanding pre-training tasks, Chen et al. (2024d)'s contradictory historical tasks and Ma et al. (2025)'s trial-and-demonstration framework. Additionally, ***scaffolding theory*** (Park and Reuter-Lorenz, 2009), which ***emphasizes gradually challenging interactions***, informs maintaining coherent learning trajectories through Borges et al. (2024)'s structured feedback loops and Sonkar et al. (2023)'s dynamic task complexity.

**Knowledge Acquisition** Semantic coherence during pre-training draw insights from ***top-down and bottom-up perception*** (Gregory, 1997), which ***frames cognition as interaction between conceptual frameworks and detailed data***. Top-down processing is leveraged to prioritize semantic processing before syntactic details (Rawte et al., 2022) and to generate test cases (Zhang et al., 2024b). Meanwhile, to enhance perception modeling, Pang et al. (2023) fuses bottom-up encoding with top-down corrections, and Nikolaus and Fourtassi (2021) models production-based learning. Introducing ***working memory theory*** (Baddeley and Hitch, 1974) that proposes ***a short-term system for temporarily holding information***, Mita et al. (2025) simulates critical period dynamics with growing memory capacity to enhance performance.

## 4 Post-Training and Alignment

With foundational knowledge acquired in pre-training, post-training refines LLMs from general proficiency to task-specific behavior. We explore

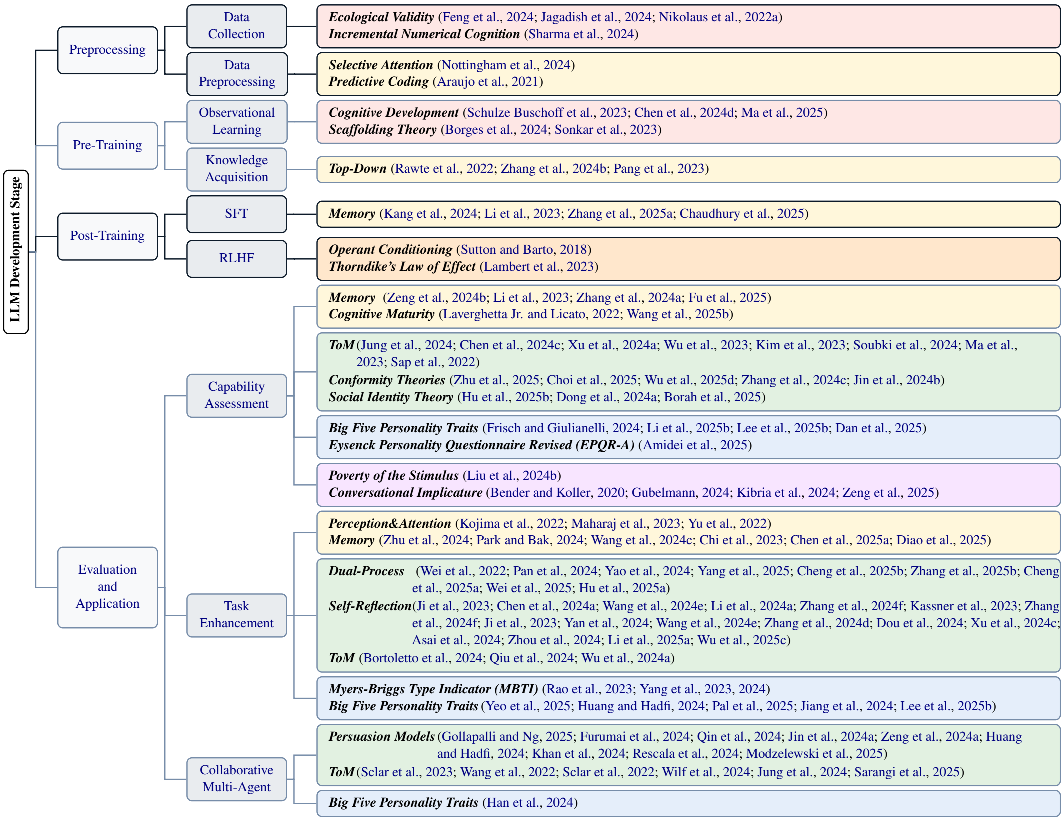


Figure 1: Our structured survey of how psychological theories apply across the main stages of LLM development. Colors indicate six distinct psychology areas: red for ***Developmental Psychology*** ; orange for ***Behavioral Psychology*** ; yellow for ***Cognitive Psychology*** ; green for ***Social Psychology*** ; blue for ***Personality Psychology*** ; purple for ***Psycholinguistics*** .

how psychology guide post-training for context-aware, interpretable, and human-aligned outcomes.

**Supervised Fine-Tuning (SFT)** In SFT, works that draw on psychological insights focus on retaining and learn contextual information. Building on ***working memory theory***, Kang et al. (2024) adds working memory module to retain short-term information, while Li et al. (2023) dynamically balances memory with contexts to improve robustness. Drawing from ***episodic memory***, ***the ability to retrieve specific experiences with contexts***(Tulving et al., 1972), Zhang et al. (2025a) enable LLMs to learn from episodic experiences for improved planning, while Chaudhury et al. (2025) introduce episodic attention for processing long contexts.

**Reinforcement Learning from Human Feedback (RLHF)** A classic behavioral theory, the ***Operant Conditioning theory*** posits that ***behaviors are systematically strengthened or weakened by the consequences (rewards or punishments) that immediately follow them*** (Thorndike, 1898; Skinner, 1957). The principles of reinforcement learning align closely with this psychological framework, particularly in the post-training phase of LLM development, where RLHF explicitly operationalizes ***Operant Conditioning theory*** to align model behaviors with human values and preferences. Through repeated feedback, the model gradually adapts to favor outputs that yield higher reward signals—a process akin, in a loose analogy, to ***Thorndike's Law of Effect***, which describes how ***behaviors followed by satisfying outcomes tend to recur***. While the underlying mechanism is driven by reward optimization algorithms rather than psychological intent, the conceptual resemblance highlights how reinforcement strategies can shape model outputs (Lambert et al., 2023). During RLHF, the model generates responses, and a learned reward function $R(x)$ assigns scores to outputs $x$, guiding subsequent policy updates. For instance, (Ouyang et al., 2022) train InstructGPT using Proximal Policy Optimization (Schulman et al., 2017), rewarding responses preferred by humans and penalizing less desirable ones. Foundational frameworks (Christiano et al., 2017; Sutton and Barto, 2018; Stiennon et al., 2022) established methods for explicitly translating human judgments into reward signals, operationalizing the insights of ***Operant Conditioning***. More recent work incorporates human cognitive biases (Siththaranjan et al., 2024) and personalizes reward functions for individual values (Poddar et al., 2024). These developments illustrate how ***Operant Conditioning*** remains central to aligning LLMs with nuanced human values. While our survey focuses on psychological dimensions, a technical overview of RLHF methods is provided in Appendix C.

## 5 Evaluation and Application

Psychology offers tools for both assessing and enhancing model behavior in the evaluation and application stage. We review three key areas of challenges where psychology can inform LLM development: **(1)** evaluating emergent capabilities such as reasoning, **(2)** improving task performance in domains involving human cognition, and **(3)** designing socially aware, multi-agent systems.

### 5.1 Benchmarks and Capability Assessment

Evaluating LLMs with psychologically grounded metrics offers a deeper window into their real-world viability. By mapping classic theories onto benchmarks that probe model responses under diverse, human-like scenarios, researchers can move beyond surface-level performance measures, revealing emergent model behavior and illuminating strengths, blind spots, and opportunities to refine LLM training and alignment practices.

#### 5.1.1 Social Reasoning and Intelligence

Social intelligence is vital for LLMs that navigate human contexts, enabling the interpretation of implicit cues, adaptation to social norms, and authentic interaction – defining advanced AI beyond mere text prediction. As LLMs increasingly mediate communication, their grasp of social dynamics becomes pivotal for both efficacy and safety.

Notably, ***Theory of Mind (ToM)*** offers a framework for evaluating ***how individuals understand and attribute mental states – such as beliefs, desires, and intentions – to others.*** By measuring LLMs' capacity to reason beliefs, researchers can assess core social intelligence. Recent benchmarks include *ToMBENCH* (Chen et al., 2024c), *OpenToM* (Xu et al., 2024a), *HITOM* (Wu et al., 2023), and *FANTOM* (Kim et al., 2023), probing distinct facets of ***ToM***. Extending the efforts to spoken dialogues, Soubki et al. (2024) reveal lingering gaps between LLM and human performance. Surveys (Ma et al., 2023; Sap et al., 2022) consolidate methods and underscore the challenges of robust ***ToM***-based evaluations.

Beyond individual cognition, social influence theories like ***Conformity Theories***(Asch, 2016), capture ***how group pressure shapes individual judgments***. Recent work tests LLM-based agents' collaboration and bias dynamics under these principles (Zhu et al., 2025; Choi et al., 2025; Wu et al., 2025d; Zhang et al., 2024c; Jin et al., 2024b), bridging individual and group-level cognition.

Emotion is another pillar of social intelligence. ***Ekman's Basic Emotion Theory*** (Ekman, 1992) identifies ***six universal emotions***, often used as labels, while ***Dimensional Models*** like the *Circumplex Model* conceptualize emotions along valence and arousal (Gong et al., 2024; Morrill et al., 2024). LLMs advance on emotion recognition, benefiting dialogue and sentiment tasks (Zhang et al., 2024e; Wu et al., 2024c,d; Sabour et al., 2024).

These efforts collectively demonstrate both progress and limitations in LLMs' social cognition, establishing benchmarks against which future developments can be measured.

#### 5.1.2 Language Proficiency

Recent work adopts psycholinguistic assessments, originally designed for humans, to test LLMs' language proficiency. These experiments probe a wide range of linguistic domains: morphology (Anh et al., 2024), syntax (Li and Hao, 2025; Amouyal et al., 2025; Liu et al., 2024b; Hale and Stanojević, 2024), phonology (Jang et al., 2025; Duan et al., 2025), semantics (Duan et al., 2025; Hayashi, 2025) and their interactions (Miaschi et al., 2024; Zhou et al., 2025a).

Although LLMs exhibit comparable performance to human speakers on many psycholinguistic tasks, the underlying processing mechanism they rely on may seem different from humans (Pedrotti et al., 2025; Lee et al., 2024). Human language acquisition is often characterized by the ***Poverty of the Stimulus***, ***children acquire complex grammar from relatively little input*** (Chomsky, 1980), whereas LLMs typically require developmentally implausible amounts of linguistic data to learn morphological rules. On the other hand, some evidence suggests that the learning patterns of LLMs mirror aspects of human language acquisition (Zhou et al., 2025b; Liu et al., 2024b).

Several studies have explored the pragmatic abilities of LLMs, motivated by the close link between language and broader cognitive functions in humans. ***Grice (1975)'s Theory of conversational implicature*** posits that ***utterance interpretation de-***

***pends on both literal content and surrounding context***. Researchers (Bender and Koller, 2020; Gubelmann, 2024) have contrasting perspectives on LLMs with respect to the Harnad (1990)'s ***Symbol Grounding Problem***, i.e. ***linguistic symbols must be grounded in sensorimotor interactions to be meaningful***. Failures of LLMs in pragmatic and semantic tasks (He et al., 2025; Kibria et al., 2024; Zeng et al., 2025), as well as their neuron patterns (Wu et al., 2024b), point to limitations beyond pure linguistic knowledge, which potentially parallel human higher-level cognitive processes.

#### 5.1.3 Memory and Cognitive Evaluation

Assessing memory and cognition is crucial given LLMs' limited capacity and risk of catastrophic forgetting. ***Memory*** is measured on parametric knowledge (Li et al., 2023), n-back tasks (Zhang et al., 2024a), capacity (Timkey and Linzen, 2023) and ***cognitive load*** (Fu et al., 2025; Xu et al., 2024b; Zeng et al., 2024b). Meanwhile, cognitive development is assessed through ***cognitive maturity*** (Wang et al., 2025b; Laverghetta Jr. and Licato, 2022), word acquisition (Chang and Bergen, 2022), ***subjective similarity*** (Malloy et al., 2024), reasoning strategies (Mondorf and Plank, 2024; Yuan et al., 2023; Ying et al., 2024), ***zone of proximal*** (Cui and Sachan, 2025) and ***perception*** (Jung et al., 2024).

#### 5.1.4 Personality Capability

Personality consistency examines how stably LLMs maintain traits across contexts. Frisch and Giulianelli (2024) show LLMs with asymmetric profiles vary in ***Big Five*** traits, while Amidei et al. (2025) find language switching alters GPT-4o's ***Eysenck Personality Questionnaire Revised*** traits, underscoring challenges in perserving consistency. Parallel research examines how LLMs display and control personality traits. Jiang et al. (2024) show LLMs express distinct traits labeled by human evaluators. Mao et al. (2024) reveals difficulties in alignment for ***Neuroticism***, ***Extraversion*** and ***Agreeableness***. Lee et al. (2025b); Li et al. (2025b); Dan et al. (2025) assess and improve consistency through alignment with psychometrical training data, while Hu and Collier (2024) find persona-based prompting improves annotation accuracy.

#### 5.1.5 Bias and Ethics Evaluation

Evaluating biases and ethical risks is crucial for responsible AI that avoids reinforcing harmful social patterns. As LLMs increasingly shape public discourse, thorough assessments are essential to prevent discriminatory outputs and promote equitable benefits across diverse communities. Recent work tests LLMs on gender (Oba et al., 2024; Zhao et al., 2024), broader social biases (Shin et al., 2024; Lee et al., 2023; Nozza et al., 2022), toxic content (Huang et al., 2025b; Gehman et al., 2020; Luong et al., 2024; Hui et al., 2024a), and harmful stereotypes (Shrawgi et al., 2024; Huang and Xiong, 2024; Hui et al., 2024b; Grigoreva et al., 2024), establishing benchmarks across cultures and languages. Evidence also suggests that LLMs replicate social identity biases, mirroring human tendencies toward ingroup favoritism and outgroup hostility (Borah et al., 2025; Hu et al., 2025b; Dong et al., 2024a) – patterns central to ***social identity theory***, which posits that ***group membership shapes self-concept and intergroup behavior*** (Tajfel, 1979).

### 5.2 Task Performance Enhancement

Building on the benchmarks, we review how psychological insights are used improves LLMs performance on complex reasoning and enrich dialogue, which illustrate how psychology improves capabilities and alignment across applications.

#### 5.2.1 Reasoning Enhancement

LLMs often struggle with complex reasoning: social inference (Liu et al., 2024a), logical errors (Turpin et al., 2023; McKenna et al., 2023), hallucinations (Huang et al., 2025a; Ai et al., 2024a), and multi-step planning (Wang et al., 2024a). Researchers address these issues by implementing analogous cognitive mechanisms. For instance, ***Dual-process theory***, a social cognition framework, ***distinguish between fast (System 1) and slow (System 2) reasoning*** (Kahneman, 2011), offers a blueprint for LLM improvement. Chain-of-thought prompting (Wei et al., 2022) operationalizes System 2 via intermediate steps, while DynaThink (Pan et al., 2024) dynamically selects rapid or thorough inference. Tree of Thoughts (Yao et al., 2024) further explores multiple reasoning paths concurrently. Yang et al. (2025) combine separate verifier as System 2. More recent applications includes hallucination mitigation (Cheng et al., 2025b), real-time human-AI collaboration (Zhang et al., 2025b), multi-hop QA (Cheng et al., 2025a), emotion consistency (Wei et al., 2025) and decoder-level LLMs merging (Hu et al., 2025a).

Similarly, ***Self-reflection and Meta-cognition***, ***introspection focused on the self-concept*** (Phillips, 2020; Flavell, 1979), has guided LLM enhancements in hallucination mitigation (Ji et al., 2023), translation (Chen et al., 2024a; Wang et al., 2024e), tool use (Li et al., 2025a), question-answering (Li et al., 2024a; Zhang et al., 2024f; Kassner et al., 2023), retrieval-augmented-generation(RAG) (Asai et al., 2024; Zhou et al., 2024) and math reasoning (Zhang et al., 2024f). Approaches include iterative self-assessment (Ji et al., 2023; Yan et al., 2024; Wu et al., 2025c), task decomposition (Wang et al., 2024e; Zhang et al., 2024d), self-training (Dou et al., 2024), and confidence-tuned reward functions (Xu et al., 2024c). Moreover, ***ToM*** adaptations boost LLMs' interpersonal reasoning, aiding missing knowledge (Bortoletto et al., 2024), common ground alignment (Qiu et al., 2024), and cognitive modeling (Wu et al., 2024a).

Beyond social reasoning, ***perception, attention, and memory*** support coherence and retrieval. Kojima et al. (2022) uses "think step by step" prompts for ***top-down*** reasoning. Chen et al. (2025a); Maharaj et al. (2023); Yu et al. (2022) leverages ***selective attention*** and ***working memory*** to detect hallucinations and extract relation. Zhu et al. (2024) employs recitation for retrieval, and Park and Bak (2024) introduce short/long-term memory modules. Diao et al. (2025); Wang et al. (2024c); Chi et al. (2023) improve reasoning via ***symbolic, adaptive and working memory*** structures. Lastly, ***hippocampal indexing theory*** (Teyler and DiScenna, 1986), ***viewing the hippocampus as a pointer to neocortical memory***, informs multi-step reasoning with external knowledge (Gutierrez et al., 2024) and counterfactual reasoning (Miao et al., 2024a).

#### 5.2.2 Dialogue Understanding and Generation

In dialogue understanding, personality psychology aids trait-based inferences from user interactions. NLP research has explored dynamic ways to measure personality beyond structured tests. The ***Myers–Briggs Type Indicator (MBTI)***, ***a self-report questionnaire that makes pseudo-scientific claims to categorize individuals into 16 distinct personality types***, remains popular (Rao et al., 2023; Yang et al., 2023), while PsychoGAT (Yang et al., 2024) gamifies ***MBTI***, and *PADO* (Yeo et al., 2025) adopts a ***Big Five***-based multi-agent approach. Beyond assessments, traits guide dialogue generation: Huang and Hadfi (2024) show higher agreeability improves negotiation, while Cheng et al. (2023) reveal social and racial biases in persona creation, raising representational concerns.

Dialogue generation research further incorporates personality to improve coherence, empathy, and consistency. Pal et al. (2025); Chen et al. (2025b) leveraged Reddit-based journal entries to model ***Big Five*** traits in large-scale dialogue datasets. ***Big Five***-aligned agents also improve on text based games (Lim et al., 2025) and code generation(Guo et al., 2025). Other efforts improve persona consistency without referencing explicit psychological theory (Wu et al., 2025b; Takayama et al., 2025). Similarly, personality is used to improve truthfulness, consistency, and context-aware generation, as further detailed in Appendix D. These approaches support personality alignment but lack grounding in deeper psychological theory.

### 5.3 Collaborative, Multi-Agent Frameworks

Beyond task-specific capabilities, the surge in multi-agent LLM frameworks reflects a growing emphasis on collaborative decision-making, where modeling social dynamics is crucial. Social and personality psychology theories offer insights to design agent interaction, negotiation, and consensus, guiding more socially intelligent LLM systems.

**Social Influence** ***Persuasion models*** (Petty and Cacioppo, 2012) illustrate ***how central/peripheral routes shape attitudes in collaborative settings***. Leveraging this, Gollapalli and Ng (2025) merges persuasive dialog acts with RL, Modzelewski et al. (2025) infuses persuasion knowledge into CoT, Furumai et al. (2024) combines LLM strategies and retrieval, Qin et al. (2024); Jin et al. (2024a) emphasize credibility-aware generation, and Zeng et al. (2024a) uncovers LLMs' vulnerabilities. Multi-agent research simulates personality-driven negotiation (Huang and Hadfi, 2024; Hu et al., 2025c), boosts truthfulness via structured debates (Khan et al., 2024), and curates argument-strength datasets (Rescala et al., 2024).

**Social Cognition** ***ToM*** complements social influence by enabling agents to grasp others' mental states. Some integrates belief tracking (Sclar et al., 2023) and coordination (Wang et al., 2022; Sclar et al., 2022), while others refine ***ToM*** via task decomposition and recursive simulation (Wilf et al., 2024; Jung et al., 2024; Sarangi et al., 2025).

**Role-Play and Multi-Agent Simulation** Recent work on persona-driven LLM agents focuses on simulating diverse perspectives, persona alignment, and socially intelligent interactions. Han et al. (2024) introduces ***Big Five***-based extraversion, Castricato et al. (2025) presents 1,586 synthetic personas, and Wu et al. (2025a) releases a benchmark with 40K multi-turn dialogues. Agents also model opinion dynamics (Wang et al., 2025a) and evaluate social intelligence (Chen et al., 2024b), with RoleLLM (Wang et al., 2024b), Character100 (Wang et al., 2024d), and persona-aware graph transformers (Mahajan and Shaikh, 2024) further supporting multi-party simulations. Lastly, Kumarage et al. (2025) simulate social engineering attacks with LLM agents of varied traits, highlighting how psychological traits shape user vulnerability.

# 6 Trends and Discussion

## 6.1 How Does Current LLM Development Harness Psychological Theories?

We observe psychological theories have been incorporated into LLM development in stage-specific ways, with uneven coverage across theoretical domains. Figure 1 maps this integration across stages.

In preprocessing and pretraining, **developmental psychology** is often referenced. Its emphasis on staged knowledge acquisition aligns with curriculum learning and progressive data exposure, mirroring human developmental trajectories. In post-training, especially RLHF, **behavioral psychology** ideas are most prominent. Conditioning, reinforcement schedules, and reward design are commonly used to guide model alignment with human preferences. In evaluation and application, theories from **social/personality psychology** and **psycholinguistics** are commonly cited, reflecting a focus on interaction patterns, user modeling, and linguistic variation – areas traditionally explored within these sub-fields. Their prominence in later stages aligns with their emphasis on human-centered communication. **Cognitive psychology** appears across all stages, particularly in modeling internal mechanisms such as reasoning, memory, and attention. Its breadth makes it a foundational influence.

The observed unevenness in integration reflects, perhaps a gap, but more probably a functional alignment – some domains are naturally better suited for certain stages of LLM development. Meanwhile, these trends expose under-explored opportunities, motivating the RQs that follow.

## 6.2 What Untapped Psychological Insights Could Advance LLM Development?

Although psychological theory is increasingly applied in LLM research, its use remains simplified and uneven. As shown in Tables 1, 2, and 3, many theories are under-utilized despite their potential to improve model behavior and interpretability. Below, we outline theories in four key areas that deserve greater attention in future LLM research.

**Social psychology** remains underutilized in areas like ***group dynamics*** and ***self and identity***, limiting personalization, adaptability, and inclusivity. Prompting LLMs to adopt specific social identities can reduce bias (Dong et al., 2024a) and mirror human-like ingroup favoritism (Hu et al., 2025b). Incorporating social identity frameworks could enhance user alignment in identity-sensitive contexts (Chen et al., 2020). Likewise, while bias detection is common, classic ***social influence theories*** (e.g., conformity, obedience) and ***attitude change theories*** (e.g., balance theory, cognitive dissonance) are rarely applied to interaction dynamics or bias mitigation, despite their relevance to ethical and socially adaptive behavior. Additionally, malicious actors leveraging social influence can severely undermine trust in digital spaces (Zeng et al., 2024a; Liu et al., 2025; Ai et al., 2024b), highlighting the potential of constructs like ***inoculation theory*** to proactively guard against manipulative strategies.

**Behavioral psychology** inspires RLHF, yet key concepts like ***partial reinforcement***, which improves behavior persistence (Ferster, 1966; Jensen, 1961), and ***shaping***, which supports gradual learning through successive approximations (Love et al., 2009), are overlooked. Current RLHF relies on uniform rewards, yet behavioral theory warns that flawed rewards can lead to reward hacking. Adding ***reward variability*** may reduce premature convergence and improve alignment with human intent (Dayan and Daw, 2008; Amodei et al., 2016).

**Personality Psychology** use focuses on ***Trait Theory***, overlooking ***developmental theories*** that explain how individual traits emerge, evolve, and adapt across contexts. These developmental models could enable more coherent and interpretable personality representations, offering a deeper alternative to static prompt-based personas.

**Cognitive psychology** remains underused, particularly ***Schema Theory***, which holds that ***humans store knowledge as schemas formed through repeated experience*** (Anderson and Pearson, 1984), guiding inference, memory, and learning. Recent work explores schema-inspired methods for compressing user histories and modeling knowledge activation cycles (Panagoulias et al., 2024; Xia et al., 2024). Further integration may improve long-term context handling and generalization.

### 6.3 What Debates Loom at the NLP–Psychology Intersection, and Where Next?

A recurring question is whether human psychology can be directly mapped to LLMs without distortion (Löhn et al., 2024). Below, we highlight key controversies at this boundary; see Appendix E for an extended discussion. These challenges motivate new recommendations and highlight open directions for cross-disciplinary exploration.

**Terminology Mismatches** A core tension is the mismatch between psychological terminology and their NLP usage. For example, **attention** in psychology means *selective mental focus*, but in transformers it is a token weighting mechanism without cognitive awareness (Lindsay, 2020), leading to misleading attributions of intentionality. Similarly, **memory** in psychology entails *structured encoding and recall*, whereas in LLMs it typically refers to context windows or parameters. Such anthropomorphic language is increasingly prevalent and shapes public and scholarly assumptions about LLMs, as recent studies show rising human-like descriptors (Ibrahim and Cheng, 2025). This calls for disentangling metaphor from mechanism through a precise cross-disciplinary lexicon, preventing both over-simplification and over-anthropomorphization – an underexplored but crucial research challenge.

**Theoretical Discrepancies** Beyond terminology, deeper theoretical mismatches arise when the NLP community adopts outdated or disputed concepts from psychology. For instance, ***predictive coding*** (Rao and Ballard, 1999) is used to analogize LLMs' next-token prediction, although current research emphasizes hierarchical, multi-scale brain mechanisms (Antonello and Huth, 2024; Caucheteux et al., 2023). Likewise, folk-psychological typologies like ***MBTI*** persist in LLM applications despite its criticized validity and reliability (Pittenger, 1993; McCrae and Costa Jr, 1989). (Wagner et al., 2025) positions that ***ToM*** involves first deciding depth of mentalizing and then applying reasoning accordingly, yet most works focus only on the latter. ***Working memory*** (Baddeley and Hitch, 1974) illustrates another gap: LLM 'memory' modules (Kang et al., 2024; Li et al., 2023) do not replicate human constraints, prompting questions about whether AI should emulate human cognitive limits or exceed them for performance gains. **Behavioral psychology** faces similar critiques (Miller, 2003; Flavell et al., 2022), as RLHF often focuses on reward optimization (Ouyang et al., 2022; Rafailov et al., 2023; Ramesh et al., 2024), neglecting internal states and risking reward hacking (Skalse et al., 2022; Krakovna, 2020). Broader debates remain over whether LLMs truly "understand" language or function as "stochastic parrots" (Ambridge and Blything, 2024; Park et al., 2024).

In response, we recommend refining how psychological theories are mapped into computational models, replacing outdated constructs with supported frameworks, exploring whether human-like constraints aid interpretability, and designing evaluations that track both outputs and internal states. Sustained collaboration between computational and psychological sciences is essential for robust and theory-aligned LLMs.

**Evaluation and Validity Debates** Another major debate is how we evaluate LLM "psychological" abilities – whether current tests really measure what they claim. For instance, GPT-4 solves around 75% of false-belief tasks, matching a 6-year-old's performance (Kosinski, 2024; Strachan et al., 2024); some see emergent ***ToM***-like reasoning (Kosinski, 2024), but others argue it may be pattern matching (Strachan et al., 2024), noting that minor prompt changes can derail results (Shapira et al., 2024). Similar controversies involve **personality**: some studies find stable simulated traits (Sorokovikova et al., 2024; Huang et al., 2024), while others reveal variability under different prompt conditions (Gupta et al., 2024; Shu et al., 2024), raising questions about inherent vs. mimicked personas (Tseng et al., 2024). This calls for more theory-grounded evaluation and clearer definitions, showing the need for a systematic, theory-driven framework beyond surface metrics, guiding more faithful replication of human cognition and behavior in LLMs.

## 7 Conclusions

We systematically review how psychology can ground LLM innovation in both past and future across severarl subfields. By examining how psychological theories inform each stage of LLM development, we find both meaningful connections across domains and critical points of tension, which are explored through discussion to help bridge interdisciplinary gaps. We hope this review sparks reflection, and inspires future work to continue integrating psychological perspectives into NLP.

## Limitations

Our review primarily focuses on literature within NLP, particularly in how personality is modeled, evaluated, and leveraged in LLMs. As a result, we do not extensively cover research from psychology and cognitive sciences that might offer deeper theoretical insights into human-like behaviors in AI. This limitation may exclude valuable methodologies or perspectives that could enhance personality evaluation frameworks for LLMs. We encourage future surveys to integrate findings from psychology and linguistics to bridge theoretical foundations with computational approaches, fostering a more comprehensive understanding of personality in AI systems.

While our survey advocates for a deeper integration of psychology into LLM design, we also caution against the ethical risks posed by overuse or misapplication of psychological principles. A concrete example is ***operant conditioning*** (Skinner, 1957), which describes how behavior can be shaped by consequences. Applied to LLMs – for instance, through timely, gratifying feedback to reinforce engagement – these mechanisms can be beneficial in contexts like language learning or motivation. However, reinforcement schedules such as variable ratio or interval rewards may unintentionally condition users to engage compulsively, raising the risk of manipulative design. This presents a key ethical limitation: distinguishing between genuinely supportive interactions and those that encourage excessive use is inherently difficult. To address this, we emphasize the need for transparent disclosure of reinforcement mechanisms and the establishment of clear ethical guidelines by professional communities. These safeguards are essential to ensure that psychological insights enhance user well-being without enabling exploitative practices.

## References


Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. Perils and opportunities in using large language models in psychological research. *PNAS nexus*, 3(7):pgae245.

Lin Ai, Zheng Hui, Zizhou Liu, and Julia Hirschberg. 2024a. Enhancing pre-trained generative language models with question attended span extraction on machine reading comprehension. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10046–10063, Miami, Florida, USA. Association for Computational Linguistics.

Lin Ai, Tharindu Sandaruwan Kumarage, Amrita Bhattacharjee, Zizhou Liu, Zheng Hui, Michael S. Davinroy, James Cook, Laura Cassani, Kirill Trapeznikov, Matthias Kirchner, Arslan Basharat, Anthony Hoogs, Joshua Garland, Huan Liu, and Julia Hirschberg. 2024b. Defending against social engineering attacks in the age of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12880–12902, Miami, Florida, USA. Association for Computational Linguistics.

Guilherme FCF Almeida, José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo. 2024. Exploring the psychology of llms' moral and legal reasoning. *Artificial Intelligence*, 333:104145.

Ben Ambridge and Liam Blything. 2024. Large language models are better than theoretical linguists at theoretical linguistics. *Theoretical Linguistics*, 50(1-2):33–48.

Jacopo Amidei, Jose Gregorio Ferreira De Sá, Rubén Nieto Luna, and Andreas Kaltenbrunner. 2025. Exploring the impact of language switching on personality traits in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2370–2378, Abu Dhabi, UAE. Association for Computational Linguistics.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in ai safety. *Preprint*, arXiv:1606.06565.

Samuel Joseph Amouyal, Aya Meltzer-Asscher, and Jonathan Berant. 2025. When the LM misunderstood the human chuckled: Analyzing garden path effects in humans and language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8235–8253, Vienna, Austria. Association for Computational Linguistics.

Richard C Anderson and P David Pearson. 1984. A schema-theoretic view of basic processes in reading comprehension. *Handbook of reading research*, 1:255–291.

Dang Anh, Limor Raviv, and Lukas Galke. 2024. Morphology matters: Probing the cross-linguistic morphological generalization abilities of large language models through a wug test. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.

Richard Antonello and Alexander Huth. 2024. Predictive coding or just feature discovery? an alternative account of why language models fit brain data. *Neurobiology of Language*, 5(1):64–79.

Vladimir Araujo, Andrés Villa, Marcelo Mendoza, Marie-Francine Moens, and Alvaro Soto. 2021. Augmenting BERT-style models with predictive coding to improve discourse-level representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3016–3022, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Solomon E Asch. 2016. Effects of group pressure upon the modification and distortion of judgments. In *Organizational influence processes*, pages 295–303. Routledge.

Alan D. Baddeley and Graham Hitch. 1974. Working memory. volume 8 of *Psychology of Learning and Motivation*, pages 47–89. Academic Press.

Simon Baron-Cohen. 2012. *The science of evil: On empathy and the origins of cruelty*. Basic books.

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Daryl J Bem. 1972. Self-perception theory. *Advances in experimental social psychology*, 6.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Angana Borah, Marwa Houalla, and Rada Mihalcea. 2025. Mind the (belief) gap: Group identity in the world of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18441–18463, Vienna, Austria. Association for Computational Linguistics.

Beatriz Borges, Niket Tandon, Tanja Käser, and Antoine Bosselut. 2024. Let me teach you: Pedagogical foundations of feedback for language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12082–12104, Miami, Florida, USA. Association for Computational Linguistics.

Matteo Bortoletto, Constantin Ruhdorfer, Adnen Abdessaied, Lei Shi, and Andreas Bulling. 2024. Limits of theory of mind modelling in dialogue-based collaborative plan acquisition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4856–4871, Bangkok, Thailand. Association for Computational Linguistics.

Meng Cao, Lei Shu, Lei Yu, Yun Zhu, Nevan Wichers, Yinxiao Liu, and Lei Meng. 2024. Enhancing reinforcement learning with dense rewards from language model critic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9119–9138, Miami, Florida, USA. Association for Computational Linguistics.

D.S. Cartwright. 1979. *Theories and Models of Personality*. W. C. Brown Company.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2025. PERSONA: A reproducible testbed for pluralistic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11348–11368, Abu Dhabi, UAE. Association for Computational Linguistics.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441.

Tyler A. Chang and Benjamin K. Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Subhajit Chaudhury, Payel Das, Sarathkrishna Swaminathan, Georgios Kollias, Elliot Nelson, Khushbu Pahwa, Tejaswini Pedapati, Igor Melnyk, and Matthew Riemer. 2025. EpMAN: Episodic memory AttentioN for generalizing to longer contexts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11696–11708, Vienna, Austria. Association for Computational Linguistics.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024a. DUAL-REFLECT: Enhancing large language models for reflective translation through dual learning feedback mechanisms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 693–704, Bangkok, Thailand. Association for Computational Linguistics.

Guanyi Chen, Yinhe Zheng, and Yupei Du. 2020. Listener's social identity matters in personalised response generation. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 205–215, Dublin, Ireland. Association for Computational Linguistics.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024b. SocialBench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126, Bangkok, Thailand. Association for Computational Linguistics.

Mingda Chen, Yang Li, Karthik Padthe, Rulin Shao, Alicia Yi Sun, Luke Zettlemoyer, Gargi Ghosh, and Wen-tau Yih. 2025a. Improving factuality with explicit working memory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11199–11213, Vienna, Austria. Association for Computational Linguistics.

Run Chen, Jun Shin, and Julia Hirschberg. 2025b. Synthempathy: A scalable empathy corpus generated using llms without any crowdsourcing. *Preprint*, arXiv:2502.17857.

Yuxuan Chen, Wei Wei, Shixuan Fan, Kaihe Xu, and Dangyang Chen. 2025c. CoMIF: Modeling of complex multiple interaction factors for conversation generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7355–7366, Abu Dhabi, UAE. Association for Computational Linguistics.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024c. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.

Ziyang Chen, Dongfang Li, Xiang Zhao, Baotian Hu, and Min Zhang. 2024d. Temporal knowledge question answering via abstract reasoning induction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4872–4889, Bangkok, Thailand. Association for Computational Linguistics.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Rong Cheng, Jinyi Liu, Yan Zheng, Fei Ni, Jiazhen Du, Hangyu Mao, Fuzheng Zhang, Bo Wang, and Jianye Hao. 2025a. DualRAG: A dual-process approach to integrate reasoning and retrieval for multi-hop question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31877–31899, Vienna, Austria. Association for Computational Linguistics.

Xiaoxue Cheng, Junyi Li, Xin Zhao, and Ji-Rong Wen. 2025b. Think more, hallucinate less: Mitigating hallucinations via dual process of fast and slow thinking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7979–7990, Vienna, Austria. Association for Computational Linguistics.

Ta-Chung Chi, Ting-Han Fan, Alexander Rudnicky, and Peter Ramadge. 2023. Transformer working memory enables regular language reasoning and natural language length extrapolation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5972–5984, Singapore. Association for Computational Linguistics.

Min Choi, Keonwoo Kim, Sungwon Chae, and Sangyeop Baek. 2025. An empirical study of group conformity in multi-agent systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5123–5139, Vienna, Austria. Association for Computational Linguistics.

N. Chomsky. 1957. *Syntactic Structures*. Janua linguarum (Mouton, Paris).: Series Minor. Mouton.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*, 50 edition. The MIT Press.

Noam Chomsky. 1980. Rules and representations. *Behavioral and Brain Sciences*, 3(1):1–15.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Peng Cui and Mrinmaya Sachan. 2025. Investigating the zone of proximal development of language models for in-context learning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6470–6483, Albuquerque, New Mexico. Association for Computational Linguistics.

Yuhao Dan, Jie Zhou, Qin Chen, Junfeng Tian, and Liang He. 2025. P-react: Synthesizing topic-adaptive reactions of personality traits via mixture of specialized LoRA experts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6342–6362, Vienna, Austria. Association for Computational Linguistics.

Peter Dayan and Nathaniel D Daw. 2008. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4):429–453.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Xingjian Diao, Chunhui Zhang, Weiyi Wu, Zhongyu Ouyang, Peijun Qing, Ming Cheng, Soroush Vosoughi, and Jiang Gui. 2025. Temporal working memory: Query-guided segment refinement for enhanced multimodal understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3393–3409, Albuquerque, New Mexico. Association for Computational Linguistics.

Wenchao Dong, Assem Zhunis, Dongyoung Jeong, Hyojin Chin, Jiyoung Han, and Meeyoung Cha. 2024a. Persona setting pitfall: Persistent outgroup biases in large language models arising from social identity adoption. *arXiv preprint arXiv:2409.03843*.

Xiaofei Dong, Xueqiang Zhang, Weixin Bu, Dan Zhang, and Feng Cao. 2024b. A survey of llm-based agents: Theories, technologies, applications and suggestions. In *2024 3rd International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC)*, pages 407–413. IEEE.

Zi-Yi Dou, Cheng-Fu Yang, Xueqing Wu, Kai-Wei Chang, and Nanyun Peng. 2024. Re-ReST: Reflection-reinforced self-training for language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15394–15411, Miami, Florida, USA. Association for Computational Linguistics.

Xufeng Duan, Xinyu Zhou, Bei Xiao, and Zhenguang Cai. 2025. Unveiling language competence neurons: A psycholinguistic approach to model interpretability. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10148–10157, Abu Dhabi, UAE. Association for Computational Linguistics.

Paul Ekman. 1992. Are there basic emotions?

HJ Eysenck and SBG Eysenck. 1984. Eysenck personality questionnaire-revised.

Steven Y. Feng, Noah Goodman, and Michael Frank. 2024. Is child-directed speech effective training data for language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22055–22071, Miami, Florida, USA. Association for Computational Linguistics.

Fernanda Ferreira and Nikole D Patson. 2007. The ‘good enough’approach to language comprehension. *Language and linguistics compass*, 1(1-2):71–83.

Ch B Ferster. 1966. Animal behavior and mental illness. *The Psychological Record*, 16(3):345–356.

Filippo Ficarra, Ryan Cotterell, and Alex Warstadt. 2025. A distributional perspective on word learning in neural language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11184–11207, Albuquerque, New Mexico. Association for Computational Linguistics.

Susan T Tufts Fiske and Shelley E Taylor. 2020. Social cognition: From brains to culture.

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

Steven W Flavell, Nadine Gogolla, Matthew Lovett-Barron, and Moriel Zelikowsky. 2022. The emergence and influence of internal states. *Neuron*, 110(16):2545–2570.

Lyn Frazier and Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14(2):178–210.

Ivar Frisch and Mario Giulianelli. 2024. LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 102–111, St. Julians, Malta. Association for Computational Linguistics.

Qihang Fu, Yongbin Qin, Ruizhang Huang, Yanping Chen, Yulin Zhou, and Lintao Long. 2025. Exclusion of thought: Mitigating cognitive load in large language models for enhanced reasoning in multiple-choice tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21673–21686, Vienna, Austria. Association for Computational Linguistics.

Kazuaki Furumai, Roberto Legaspi, Julio Cesar Vizcarra Romero, Yudai Yamazaki, Yasutaka Nishimura, Sina Semnani, Kazushi Ikeda, Weiyan Shi, and Monica Lam. 2024. Zero-shot persuasive chatbots with LLM-generated strategies and information retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11224–11249, Miami, Florida, USA. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Matthew Goldrick and Brenda Rapp. 2007. Lexical and post-lexical phonological representations in spoken production. *Cognition*, 102(2):219–260.

Sujatha Das Gollapalli and See-Kiong Ng. 2025. PIRsuader: A persuasive chatbot for mitigating psychological insulin resistance in type-2 diabetic patients. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5997–6013, Abu Dhabi, UAE. Association for Computational Linguistics.

Ziwei Gong, Muyin Yao, Xinyi Hu, Xiaoning Zhu, and Julia Hirschberg. 2024. A mapping on current classifying categories of emotions used in multimodal models for emotion recognition. In *Proceedings of The*

18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 19–28, St. Julians, Malta. Association for Computational Linguistics.

Richard L Gregory. 1997. Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1358):1121–1127.

H. Paul Grice. 1975. Logic and conversation. In Donald Davidson, editor, *The logic of grammar*, pages 64–75. Dickenson Pub. Co.

Veronika Grigoreva, Anastasiia Ivanova, Ilseyar Alimova, and Ekaterina Artemova. 2024. RuBia: A Russian language bias detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14227–14239, Torino, Italia. ELRA and ICCL.

Reto Gubelmann. 2024. Pragmatic norms are all you need – why the symbol grounding problem does not apply to LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11663–11678, Miami, Florida, USA. Association for Computational Linguistics.

Yaoqi Guo, Zhenpeng Chen, Jie M. Zhang, Yang Liu, and Yun Ma. 2025. Personality-guided code generation using large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1068–1080, Vienna, Austria. Association for Computational Linguistics.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2024. Self-assessment tests are unreliable measures of LLM personality. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 301–314, Miami, Florida, US. Association for Computational Linguistics.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie C. Y. Chan, Andrew Lampinen, Jane X. Wang, Zeynep Akata, and Eric Schulz. 2023. Machine Psychology. *arXiv e-prints*, arXiv:2303.13988.

John T. Hale and Miloš Stanojević. 2024. Do LLMs learn a true syntactic universal? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17106–17119, Miami, Florida, USA. Association for Computational Linguistics.

Ji-Eun Han, Jun-Seok Koh, Hyeon-Tae Seo, Du-Seong Chang, and Kyung-Ah Sohn. 2024. PSYDIAL: Personality-based synthetic dialogue generation using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13321–13331, Torino, Italia. ELRA and ICCL.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346.

Yoshihiko Hayashi. 2025. Evaluating LLMs' capability to identify lexical semantic equivalence: Probing with the word-in-context task. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6985–6998, Abu Dhabi, UAE. Association for Computational Linguistics.

Linyang He, Ercong Nie, Helmut Schmid, Hinrich Schuetze, Nima Mesgarani, and Jonathan Brennan. 2025. Large language models as neurolinguistic subjects: Discrepancy between performance and competence. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19284–19302, Vienna, Austria. Association for Computational Linguistics.

Fritz Heider. 1946. Attitudes and cognitive organization. *The Journal of psychology*, 21(1):107–112.

Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W. Bradley Knox, and Dorsa Sadigh. 2024. Contrastive preference learning: Learning from human feedback without reinforcement learning. In *The Twelfth International Conference on Learning Representations*.

Eghbal A Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. 2024. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, 5(1):43–63.

Bodun Hu, Shuozhe Li, Saurabh Agarwal, Myungjin Lee, Akshay Jajoo, Jiamin Li, Le Xu, Geon-Woo Kim, Donghyun Kim, Hong Xu, Amy Zhang, and Aditya Akella. 2025a. StitchLLM: Serving LLMs, one block at a time. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26887–26903, Vienna, Austria. Association for Computational Linguistics.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2025b. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75.

Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 2025c. Debate-to-write: A persona-driven multi-agent framework for diverse argument generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4689–4703, Abu Dhabi, UAE. Association for Computational Linguistics.

Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael Lyu. 2024. On the reliability of psychological scales on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6152–6173, Miami, Florida, USA. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Yin Jou Huang and Rafik Hadfi. 2024. How personality traits influence negotiation outcomes? a simulation based on large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10336–10351, Miami, Florida, USA. Association for Computational Linguistics.

Yufei Huang and Deyi Xiong. 2024. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2917–2929, Torino, Italia. ELRA and ICCL.

Yuyi Huang, Runzhe Zhan, Derek F. Wong, Lidia S. Chao, and Ailin Tao. 2025b. Intrinsic model weaknesses: How priming attacks unveil vulnerabilities in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1405–1425, Albuquerque, New Mexico. Association for Computational Linguistics.

Zheng Hui, Zhaoxiao Guo, Hang Zhao, Juanyong Duan, Lin Ai, Yinheng Li, Julia Hirschberg, and Congrui Huang. 2024a. Can open-source llms enhance data augmentation for toxic detection?: An experimental study. *arXiv preprint arXiv:2411.15175*.

Zheng Hui, Zhaoxiao Guo, Hang Zhao, Juanyong Duan, and Congrui Huang. 2024b. Toxicraft: A novel framework for synthetic generation of harmful information. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16632–16647.

Lujain Ibrahim and Myra Cheng. 2025. Thinking beyond the anthropomorphic paradigm benefits llm research. *arXiv preprint arXiv:2502.09192*.

Gary M Ingersoll, Donald P Orr, Alison J Herrold, and Michael P Golden. 1986. Cognitive maturity and self-management among adolescents with insulin-dependent diabetes mellitus. *The Journal of pediatrics*, 108(4):620–623.

Akshay K. Jagadish, Julian Coda-Forno, Mirko Thalmann, Eric Schulz, and Marcel Binz. 2024. Human-like category learning by injecting ecological priors from large language models into neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Dongjun Jang, Youngchae Ahn, and Hyopil Shin. 2025. P-CoT: A pedagogically-motivated participatory chain-of-thought prompting for phonological reasoning in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21958–21979, Vienna, Austria. Association for Computational Linguistics.

Irving L Janis. 1972. Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes.

Glen D. Jensen. 1961. Partial reinforcement effects (pres) and inverse pres determined by position of a nonrewarded block of responses. *Journal of Experimental Psychology*, 62(5):461.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024a. Persuading across diverse domains: a dataset and persuasion large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706, Bangkok, Thailand. Association for Computational Linguistics.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024b. AgentReview: Exploring peer review dynamics with LLM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1208–1226, Miami, Florida, USA. Association for Computational Linguistics.

Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, and He He. 2024. Personas as a way to model truthfulness in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6346–6359, Miami, Florida, USA. Association for Computational Linguistics.

Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, and Hyunwoo Kim. 2024. Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19794–19809, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux*.

Jikun Kang, Romain Laroche, Xingdi Yuan, Adam Trischler, Xue Liu, and Jie Fu. 2024. Think before you act: Decision transformers with working memory. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 23001–23021. PMLR.

Nora Kassner, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schuetze, and Peter Clark. 2023. Language models with rationality. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14190–14201, Singapore. Association for Computational Linguistics.

Luoma Ke, Song Tong, Peng Cheng, and Kaiping Peng. 2024. Exploring the frontiers of llms in psychological applications: A comprehensive review. *Preprint*, arXiv:2401.01519.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Raihan Kibria, Sheikh Intiser Uddin Dipta, and Muhammad Abdullah Adnan. 2024. On functional competence of LLMs for linguistic disambiguation. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 143–160, Miami, FL, USA. Association for Computational Linguistics.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore. Association for Computational Linguistics.

Jinsung Kim, Seonmin Koo, and Heuiseok Lim. 2024. PANDA: Persona attributes navigation for detecting and alleviating overuse problem in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12005–12026, Miami, Florida, USA. Association for Computational Linguistics.

Walter Kintsch. 1988. The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review*, 95(2):163.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Victoria Krakovna. 2020. Specification gaming: The flip side of ai ingenuity.

Tharindu Kumarage, Cameron Johnson, Jadie Adams, Lin Ai, Matthias Kirchner, Anthony Hoogs, Joshua Garland, Julia Hirschberg, Arslan Basharat, and Huan Liu. 2025. Personalized attacks of social engineering in multi-turn conversations–llm agents for simulation and detection. *arXiv preprint arXiv:2503.15552*.

Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. 2023. The history and risks of reinforcement learning and human feedback. *Preprint*, arXiv:2310.13595.

Bibb Latané. 1981. The psychology of social impact. *American psychologist*, 36(4):343.

Bibb Latané, Kipling Williams, and Stephen Harkins. 1979. Many hands make light the work: The causes and consequences of social loafing. *Journal of personality and social psychology*, 37(6):822.

Antonio Laverghetta Jr. and John Licato. 2022. Developmental negation processing in transformer language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 545–551, Dublin, Ireland. Association for Computational Linguistics.

Eun-Kyoung Rosa Lee, Sathvik Nair, and Naomi Feldman. 2024. A psycholinguistic evaluation of language models' sensitivity to argument roles. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3262–3274, Miami, Florida, USA. Association for Computational Linguistics.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-woo Ha. 2023. KoSBI: A dataset for mitigating social bias risks towards safer large language model applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 208–224, Toronto, Canada. Association for Computational Linguistics.

Keyeun Lee, Seo Hyeong Kim, Seolhee Lee, Jinsu Eun, Yena Ko, Hayeon Jeon, Esther Hehsun Kim, Seonghye Cho, Soeun Yang, Eun-mee Kim, and Hajin Lim. 2025a. SPeCtrum: A grounded framework for multidimensional identity representation in LLM-based agent. In *Proceedings of the 2025 Conference* *of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6971–6991, Albuquerque, New Mexico. Association for Computational Linguistics.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025b. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8397–8437, Albuquerque, New Mexico. Association for Computational Linguistics.

Willem JM Levelt, Ardi Roelofs, and Antje S Meyer. 1999. A theory of lexical access in speech production. *Behavioral and brain sciences*, 22(1):1–38.

Clayton Lewis. 2025. Artificial psychology. *Synthesis Lectures on Human-Centered Informatics*.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Qinchan Li and Sophie Hao. 2025. ERAS: Evaluating the robustness of Chinese NLP models to morphological garden path errors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3100–3111, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 2025b. BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20434–20471, Vienna, Austria. Association for Computational Linguistics.

Yanhong Li, Chenghao Yang, and Allyson Ettinger. 2024a. When hindsight is not 20/20: Testing limits on reflective thinking in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3741–3753, Mexico City, Mexico. Association for Computational Linguistics.

Yuan Li, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. 2024b. Quantifying ai psychology: A psychometrics benchmark for large language models. *Preprint*, arXiv:2406.17675.

Seungwon Lim, Seungbeen Lee, Dongjun Min, and Youngjae Yu. 2025. Persona dynamics: Unveiling the impact of persona traits on agents in text-based games. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31360–31394, Vienna, Austria. Association for Computational Linguistics.

Grace W Lindsay. 2020. Attention in psychology, neuroscience, and machine learning. *Frontiers in computational neuroscience*, 14:29.

Jiateng Liu, Lin Ai, Zizhou Liu, Payam Karisani, Zheng Hui, Yi Fung, Preslav Nakov, Julia Hirschberg, and Heng Ji. 2025. PropaInsight: Toward deeper understanding of propaganda in terms of techniques, appeals, and intent. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5607–5628, Abu Dhabi, UAE. Association for Computational Linguistics.

Xuan Liu, Jie Zhang, Haoyang Shang, Song Guo, Chengxu Yang, and Quanyan Zhu. 2024a. Exploring prosocial irrationality for llm agents: A social cognition view. *arXiv preprint arXiv:2405.14744*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. 2024b. Zhoblimp: a systematic assessment of language models with linguistic minimal pairs in chinese. *Preprint*, arXiv:2411.06096.

Lea Löhn, Niklas Kiehne, Alexander Ljapunov, and Wolf-Tilo Balke. 2024. Is machine psychology here? on requirements for using human psychological tests on large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 230–242, Tokyo, Japan. Association for Computational Linguistics.

Jessa R. Love, James E. Carr, Season M. Almason, and Anna Ingeborg Petursdottir. 2009. Early and intensive behavioral intervention for autism: A survey of clinical practices. *Research in Autism Spectrum Disorders*, 3(2):421–428.

Tinh Luong, Thanh-Thien Le, Linh Ngo, and Thien Nguyen. 2024. Realistic evaluation of toxicity in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1038–1047, Bangkok, Thailand. Association for Computational Linguistics.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory of mind in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031, Singapore. Association for Computational Linguistics.

Ziqiao Ma, Zekun Wang, and Joyce Chai. 2025. Babysit a language model from scratch: Interactive language learning by trials and demonstrations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 991–1010, Albuquerque, New Mexico. Association for Computational Linguistics.

Maryellen C MacDonald, Neal J Pearlmutter, and Mark S Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological review*, 101(4):676.

Khyati Mahajan and Samira Shaikh. 2024. Persona-aware multi-party conversation response generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12712–12723, Torino, Italia. ELRA and ICCL.

Kishan Maharaj, Ashita Saxena, Raja Kumar, Abhijit Mishra, and Pushpak Bhattacharyya. 2023. Eyes show the way: Modelling gaze behaviour for hallucination detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11424–11438, Singapore. Association for Computational Linguistics.

David R Maines. 1989. Rediscovering the social group: A self-categorization theory.

Tyler Malloy, Maria José Ferreira, Fei Fang, and Cleotilde Gonzalez. 2024. Leveraging a cognitive model to measure subjective similarity of human and GPT-4 written content. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 517–527, Miami, FL, USA. Association for Computational Linguistics.

Shengyu Mao, Xiaohan Wang, Mengru Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Ningyu Zhang. 2024. Editing personality for large language models. In *Natural Language Processing and Chinese Computing: 13th National CCF Conference, NLPCC 2024, Hangzhou, China, November 1–3, 2024, Proceedings, Part II*, page 241–254, Berlin, Heidelberg. Springer-Verlag.

Robert R McCrae and Paul T Costa Jr. 1989. Reinterpreting the myers-briggs type indicator from the perspective of the five-factor model of personality. *Journal of personality*, 57(1):17–40.

WJ McGuire. 1964. Inducing resistance to persuasion: Some contemporary approaches. *Advances in Experimental Social Psychology/Academic Press*.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2758–2774, Singapore. Association for Computational Linguistics.

Xin Miao, Yongqi Li, Shen Zhou, and Tieyun Qian. 2024a. Episodic memory retrieval from LLMs: A neuromorphic mechanism to generate commonsense counterfactuals for relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2489–2511, Bangkok, Thailand. Association for Computational Linguistics.

Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. 2024b. InfoRM: Mitigating reward hacking in RLHF via information-theoretic reward modeling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Alessio Miaschi, Felice Dell'Orletta, and Giulia Venturi. 2024. Evaluating large language models via linguistic profiling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2835–2848, Miami, Florida, USA. Association for Computational Linguistics.

Stanley Milgram. 1963. Behavioral study of obedience. *The Journal of abnormal and social psychology*, 67(4):371.

George A Miller. 2003. The cognitive revolution: a historical perspective. *Trends in cognitive sciences*, 7(3):141–144.

Masato Mita, Ryo Yoshida, and Yohei Oseki. 2025. Developmentally-plausible working memory shapes a critical period for language acquisition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9386–9399, Vienna, Austria. Association for Computational Linguistics.

Arkadiusz Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, and Giovanni Da San Martino. 2025. PCoT: Persuasion-augmented chain of thought for detecting fake news and social media disinformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24959–24983, Vienna, Austria. Association for Computational Linguistics.

Philipp Mondorf and Barbara Plank. 2024. Comparing inferential strategies of humans and large language models in deductive reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9370–9402, Bangkok, Thailand. Association for Computational Linguistics.

Todd Morrill, Zhaoyuan Deng, Yanda Chen, Amith Ananthram, Colin Wayne Leach, and Kathleen McKeown. 2024. Social orientation: A new feature for dialogue analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14995–15011, Torino, Italia. ELRA and ICCL.

Camille Morvan and Alexander O'Connor. 2017. *An analysis of Leon Festinger's a theory of cognitive dissonance*. Macat Library.

Ted Moskovitz, Aaditya K Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca Dragan, and Stephen Marcus McAleer. 2024. Confronting reward model overoptimization with constrained RLHF. In *The Twelfth International Conference on Learning Representations*.

I.B. Myers and P.B. Myers. 1995. *Gifts Differing: Understanding Personality Type*. Mobius.

Mitja Nikolaus, Afra Alishahi, and Grzegorz Chrupała. 2022a. Learning English with Peppa Pig. *Transactions of the Association for Computational Linguistics*, 10:922–936.

Mitja Nikolaus, Afra Alishahi, and Grzegorz Chrupała. 2022b. Learning English with Peppa Pig. *Transactions of the Association for Computational Linguistics*, 10:922–936.

Mitja Nikolaus and Abdellah Fourtassi. 2021. Modeling the interaction between perception-based and production-based learning in children's early acquisition of semantic knowledge. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 391–407, Online. Association for Computational Linguistics.

Kolby Nottingham, Yasaman Razeghi, Kyungmin Kim, Jb Lanier, Pierre Baldi, Roy Fox, and Sameer Singh. 2024. Selective perception: Learning concise state descriptions for language model actors. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 327–341, Mexico City, Mexico. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 68–74, virtual+Dublin. Association for Computational Linguistics.

Daisuke Oba, Masahiro Kaneko, and Danushka Bollegala. 2024. In-contextual gender bias suppression for large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1722–1742, St. Julian's, Malta. Association for Computational Linguistics.

Desmond Ong. 2024. Gpt-ology, computational models, silicon sampling: How should we think about llms in cognitive science? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Sayantan Pal, Souvik Das, and Rohini K. Srihari. 2025. Beyond discrete personas: Personality modeling through journal intensive conversations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7055–7074, Abu Dhabi, UAE. Association for Computational Linguistics.

Jiabao Pan, Yan Zhang, Chen Zhang, Zuozhu Liu, Hongwei Wang, and Haizhou Li. 2024. DynaThink: Fast or slow? a dynamic decision-making framework for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14686–14695, Miami, Florida, USA. Association for Computational Linguistics.

Dimitrios P. Panagoulias, Persephone Papatheodosiou, Anastasios Bonakis, Dimitrios Dikeos, Maria Virvou, and George A. Tsihrintzis. 2024. Memory and schema in human-generative artificial intelligence interactions. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 462–467.

Bo Pang, Erik Nijkamp, Wojciech Kryscinski, Silvio Savarese, Yingbo Zhou, and Caiming Xiong. 2023. Long document summarization with top-down and bottom-up inference. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1267–1284, Dubrovnik, Croatia. Association for Computational Linguistics.

Denise C Park and Patricia Reuter-Lorenz. 2009. The adaptive brain: aging and neurocognitive scaffolding. *Annual review of psychology*, 60(1):173–196.

Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park, and Sungeun Lee. 2024. MultiPragEval: Multilingual pragmatic evaluation of large language models. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 96–119, Miami, Florida, USA. Association for Computational Linguistics.

Sangjun Park and JinYeong Bak. 2024. Memoria: resolving fateful forgetting problem through human-inspired memory architecture. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

P Ivan Pavlov. 1927. Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. *Annals of neurosciences*, 17(3):136.

Andrea Pedrotti, Giulia Rambelli, Caterina Villani, and Marianna Bolognesi. 2025. How humans and LLMs organize conceptual knowledge: Exploring subordinate categories in Italian. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4464–4482, Vienna, Austria. Association for Computational Linguistics.

Max Pellert, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826. PMID: 38165766.

Richard E Petty and John T Cacioppo. 2012. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer Science & Business Media.

Ann G. Phillips. 2020. *Self-Reflection*, pages 4791–4794. Springer International Publishing, Cham.

Jean Piaget. 1976. *Piaget's theory*. Springer.

Jean Piaget. 2013. *Child's Conception of Number: Selected Works vol 2*. Routledge.

W.D. Pierce and C.D. Cheney. 2008. *Behavior Analysis and Learning*. Psychology Press.

David J Pittenger. 1993. Measuring the mbti… and coming up short. *Journal of Career Planning and Employment*, 54(1):48–52.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

David Premack. 1959. Toward empirical behavior laws: I. positive reinforcement. *Psychological review*, 66(4):219.

Peixin Qin, Chen Huang, Yang Deng, Wenqiang Lei, and Tat-Seng Chua. 2024. Beyond persuasion: Towards conversational recommender system with credible explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4264–4282, Miami, Florida, USA. Association for Computational Linguistics.

Shuwen Qiu, Mingdian Liu, Hengli Li, Song-Chun Zhu, and Zilong Zheng. 2024. MindDial: Enhancing conversational agents with theory-of-mind for common ground alignment and negotiation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 746–759, Kyoto, Japan. Association for Computational Linguistics.

Youzhi Qu, Penghui Du, Wenxin Che, Chen Wei, Chi Zhang, Wanli Ouyang, Yatao Bian, Feiyang Xu, Bin Hu, Kai Du, et al. 2024. Promoting interactions between cognitive science and large language models. *The Innovation*, 5(2).

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. 2024. Group robust preference optimization in reward-free rlhf. In *Advances in Neural Information Processing Systems*, volume 37, pages 37100–37137. Curran Associates, Inc.

Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can ChatGPT assess human personalities? a general evaluation framework. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1184–1194, Singapore. Association for Computational Linguistics.

Rajesh PN Rao and Dana H Ballard. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87.

Vipula Rawte, Megha Chakraborty, Kaushik Roy, Manas Gaur, Keyur Faldu, Prashant Kikani, Hemang Akbari, and Amit P. Sheth. 2022. TDLR: Top semantic-down syntactic language representation. In *NeurIPS '22 Workshop on All Things Attention: Bridging Different Perspectives on Attention*.

Paula Rescala, Manoel Horta Ribeiro, Tiancheng Hu, and Robert West. 2024. Can language models recognize convincing arguments? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8826–8837, Miami, Florida, USA. Association for Computational Linguistics.

Maor Reuben, Ortal Slobodin, Idan-Chaim Cohen, Aviad Elyashar, Orna Braun-Lewensohn, Odeya Cohen, and Rami Puzis. 2025. Assessment and manipulation of latent constructs in pre-trained language models using psychometric scales. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2433–2444, Vienna, Austria. Association for Computational Linguistics.

Sonia Roccas, Lilach Sagiv, Shalom H. Schwartz, and Ariel Knafo. 2002. The big five personality factors and personal values. *Personality and Social Psychology Bulletin*, 28(6):789–801.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting* of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sneheel Sarangi, Maha Elgarf, and Hanan Salam. 2025. Decompose-ToM: Enhancing theory of mind reasoning in large language models through simulation and task decomposition. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10228–10241, Abu Dhabi, UAE. Association for Computational Linguistics.

Giuseppe Sartori and Graziella Orrù. 2023. Language models and psychological sciences. *Frontiers in Psychology*, 14:1279317.

David E Scharff, WRD Fairbairn, and Ellinor Fairbairn Birtles. 2013. *Psychoanalytic studies of the personality*. Routledge.

Mark A Schmuckler. 2001. What is ecological validity? a dimensional analysis. *Infancy*, 2(4):419–436.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Wolfram Schultz. 1998. Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, 80(1):1–27.

Luca M. Schulze Buschoff, Eric Schulz, and Marcel Binz. 2023. The acquisition of physical knowledge in generative neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980, Toronto, Canada. Association for Computational Linguistics.

Melanie Sclar, Graham Neubig, and Yonatan Bisk. 2022. Symmetric machine theory of mind. In *International Conference on Machine Learning*, pages 19450–19466. PMLR.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta. Association for Computational Linguistics.

Mandar Sharma, Rutuja Taware, Pravesh Koirala, Nikhil Muralidhar, and Naren Ramakrishnan. 2024. Laying anchors: Semantically priming numerals in language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2653–2660, Mexico City, Mexico. Association for Computational Linguistics.

Lingfeng Shen, Sihao Chen, Linfeng Song, Lifeng Jin, Baolin Peng, Haitao Mi, Daniel Khashabi, and Dong Yu. 2024. The trickle-down impact of reward inconsistency on RLHF. In *The Twelfth International Conference on Learning Representations*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong Park. 2024. Ask LLMs directly, "what shapes your bias?": Measuring social bias in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16122–16143, Bangkok, Thailand. Association for Computational Linguistics.

Hari Shrawgi, Prasanjit Rath, Tushar Singhal, and Sandipan Dandapat. 2024. Uncovering stereotypes in large language models: A task complexity-based approach. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1841–1857, St. Julian's, Malta. Association for Computational Linguistics.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281, Mexico City, Mexico. Association for Computational Linguistics.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2024. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *The Twelfth International Conference on Learning Representations*.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471.

Burrhus F Skinner. 1938. Operant behavior. *American psychologist*, 18(8):503.

Burrhus Frederic Skinner. 1957. *The behavior of organisms: An experimental analysis*. BF Skinner Foundation.

Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. CLASS: A design framework for building intelligent tutoring systems based on learning science principles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1941–1961, Singapore. Association for Computational Linguistics.

Aleksandra Sorokovikova, Sharwin Rezagholi, Natalia Fedorova, and Ivan P. Yamshchikov. 2024. LLMs simulate big5 personality traits: Further evidence. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 83–87, St. Julians, Malta. Association for Computational Linguistics.

Adil Soubki, John Murzaku, Arash Yousefi Jordehi, Peter Zeng, Magdalena Markowska, Seyed Abolghasem Mirroshandel, and Owen Rambow. 2024. Views are my own, but also yours: Benchmarking theory of mind using common ground. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14815–14823, Bangkok, Thailand. Association for Computational Linguistics.

Rose M Spielman, William J Jenkins, and Marilyn D Lovett. 2024. *Psychology 2e*.

CM Steele. 1988. The psychology of self-affirmation: Sustaining the integrity of the self. *Advances in experimental social psychology*, 21.

P Stefaroi. 2015. Humanistic personology, a humanistic-ontological theory of the person & personality: applications in therapy. *Social work, education, management, and art (theatre). Charleston SC, USA [United States of America]: Create Space*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *Preprint*, arXiv:2009.01325.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Richard S Sutton and Andrew G Barto. 2018. Reinforcement learning: an introduction, 2nd edn. adaptive computation and machine learning.

Henri Tajfel. 1979. An integrative theory of intergroup conflict. *The social psychology of intergroup relations/Brooks/Cole*.

Junya Takayama, Masaya Ohagi, Tomoya Mizumoto, and Katsumasa Yoshikawa. 2025. Persona-consistent dialogue generation via pseudo preference tuning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5507–5514, Abu Dhabi, UAE. Association for Computational Linguistics.

Timothy J Teyler and Pascal DiScenna. 1986. The hippocampal memory indexing theory. *Behavioral neuroscience*, 100(2):147.

Edward L Thorndike. 1898. Animal intelligence.

Edward L Thorndike. 1927. The law of effect. *The American journal of psychology*, 39(1/4):212–222.

William Timkey and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8705–8720, Singapore. Association for Computational Linguistics.

Michael Tomasello. 2005. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

Anne M Treisman. 1969. Strategies and models of selective attention. *Psychological review*, 76(3):282.

Jen tse Huang, Wenxuan Wang, Eric John Li, Man Ho LAM, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2024. On the humanity of conversational AI: Evaluating the psychological portrayal of LLMs. In *The Twelfth International Conference on Learning Representations*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

Endel Tulving et al. 1972. Episodic and semantic memory. *Organization of memory*, 1(381-403):1.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Eitan Wagner, Nitay Alon, Joseph M Barnby, and Omri Abend. 2025. Mind your theory: Theory of mind goes deeper than reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26658–26668, Vienna, Austria. Association for Computational Linguistics.

Chaojie Wang, Yanchen Deng, Zhiyi Lyu, Liang Zeng, Jujie He, Shuicheng Yan, and Bo An. 2024a. Q*: Improving multi-step reasoning for llms with deliberative planning. *arXiv preprint arXiv:2406.14283*.

Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. 2025a. Decoding echo chambers: LLM-powered simulations revealing polarization in social networks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3913–3923, Abu Dhabi, UAE. Association for Computational Linguistics.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024b. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. 2024c. Symbolic working memory enhances language models for complex rule application. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17583–17604, Miami, Florida, USA. Association for Computational Linguistics.

Xi Wang, Hongliang Dai, Shen Gao, and Piji Li. 2024d. Characteristic AI agents via large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3016–3027, Torino, Italia. ELRA and ICCL.

Xinglin Wang, Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Boyuan Pan, Heda Wang, Yao Hu, and Kan Li. 2025b. CogLM: Tracking cognitive development of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 73–87, Albuquerque, New Mexico. Association for Computational Linguistics.

Yuanfei Wang, Fangwei Zhong, Jing Xu, and Yizhou Wang. 2022. Tom2c: Target-oriented multi-agent communication and cooperation with theory of mind. In *International Conference on Learning Representations*.

Yutong Wang, Jiali Zeng, Xuebo Liu, Fandong Meng, Jie Zhou, and Min Zhang. 2024e. TasTe: Teaching large language models to translate through self-reflection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6144–6158, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yangbo Wei, Zhen Huang, Fangzhou Zhao, Qi Feng, and Wei W. Xing. 2025. MECoT: Markov emotional chain-of-thought for personality-consistent role-playing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8297–8314, Vienna, Austria. Association for Computational Linguistics.

James V Wertsch. 1988. *Vygotsky and the social formation of mind*. Harvard university press.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308, Bangkok, Thailand. Association for Computational Linguistics.

Bowen Wu, Kaili Sun, Ziwei Bai, Ying Li, and Baoxun Wang. 2025a. RAIDEN benchmark: Evaluating role-playing conversational agents with measurement-driven custom dialogues. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11086–11106, Abu Dhabi, UAE. Association for Computational Linguistics.

Jiaqiang Wu, Xuandong Huang, Zhouan Zhu, and Shangfei Wang. 2025b. From traits to empathy: Personality-aware multimodal empathetic response generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8925–8938, Abu Dhabi, UAE. Association for Computational Linguistics.

Jincenzi Wu, Zhuang Chen, Jiawen Deng, Sahand Sabour, Helen Meng, and Minlie Huang. 2024a. COKE: A cognitive knowledge graph for machine theory of mind. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15984–16007, Bangkok, Thailand. Association for Computational Linguistics.

Meng-Chen Wu, Md Mosharaf Hossain, Tess Wood, Shayan Ali Akbar, Si-Chi Chin, and Erwin Cornejo. 2025c. SEEval: Advancing LLM text evaluation efficiency and accuracy through self-explanation prompting. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7357–7368, Albuquerque, New Mexico. Association for Computational Linguistics.

Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. 2024b. Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language*

*Processing*, pages 22583–22599, Miami, Florida, USA. Association for Computational Linguistics.

Weidong Wu, Qinlin Zhao, Hao Chen, Lexin Zhou, Defu Lian, and Hong Xie. 2025d. Exploring the choice behavior of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5194–5214, Vienna, Austria. Association for Computational Linguistics.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore. Association for Computational Linguistics.

Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. 2024c. Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances. *Preprint*, arXiv:2407.21315.

Zehui Wu, Ziwei Gong, Jaywon Koo, and Julia Hirschberg. 2024d. Multimodal multi-loss fusion network for sentiment analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3588–3602, Mexico City, Mexico. Association for Computational Linguistics.

Long Xia, Wenqi Shen, Weiguo Fan, and G. Alan Wang and. 2024. Knowledge-aware learning framework based on schema theory to complement large learning models. *Journal of Management Information Systems*, 41(2):453–486.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024a. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand. Association for Computational Linguistics.

Nan Xu, Fei Wang, Ben Zhou, Bangzheng Li, Chaowei Xiao, and Muhao Chen. 2024b. Cognitive overload: Jailbreaking large language models with overloaded logical thinking. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3526–3548, Mexico City, Mexico. Association for Computational Linguistics.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024c. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5985–5998, Miami, Florida, USA. Association for Computational Linguistics.

Hanqi Yan, Qinglin Zhu, Xinyu Wang, Lin Gui, and Yulan He. 2024. Mirror: Multiple-perspective self-reflection method for knowledge-rich reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7086–7103, Bangkok, Thailand. Association for Computational Linguistics.

Cheng Yang, Chufan Shi, Siheng Li, Bo Shui, Yujiu Yang, and Wai Lam. 2025. LLM2: Let large language models harness system 2 reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 168–177, Albuquerque, New Mexico. Association for Computational Linguistics.

Qisen Yang, Zekun Wang, Honghui Chen, Shenzhi Wang, Yifan Pu, Xin Gao, Wenhao Huang, Shiji Song, and Gao Huang. 2024. PsychoGAT: A novel psychological measurement paradigm through interactive fiction games with LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14470–14505, Bangkok, Thailand. Association for Computational Linguistics.

Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. 2023. PsyCoT: Psychological questionnaire as powerful chain-of-thought for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3305–3320, Singapore. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025. SweetieChat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4646–4669, Abu Dhabi, UAE. Association for Computational Linguistics.

Haein Yeo, Taehyeong Noh, Seungwan Jin, and Kyungsik Han. 2025. PADO: Personality-induced multi-agents for detecting OCEAN in human-generated texts. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5719–5736, Abu Dhabi, UAE. Association for Computational Linguistics.

Jiahao Ying, Yixin Cao, Kai Xiong, Long Cui, Yidong He, and Yongbin Liu. 2024. Intuitive or dependent? investigating LLMs' behavior style to conflicting prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4221–4246, Bangkok, Thailand. Association for Computational Linguistics.

Jiaxin Yu, Deqing Yang, and Shuyu Tian. 2022. Relation-specific attentions over entity mentions for enhanced document-level relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1523–1529, Seattle, United States. Association for Computational Linguistics.

Siyu Yuan, Jiangjie Chen, Xuyang Ge, Yanghua Xiao, and Deqing Yang. 2023. Beneath surface similarity: Large language models make reasonable scientific analogies after structure abduction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2446–2460, Singapore. Association for Computational Linguistics.

Robert B Zajonc. 1965. Social facilitation: A solution is suggested for an old unresolved social psychological problem. *Science*, 149(3681):269–274.

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. 2025. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10602–10617, Abu Dhabi, UAE. Association for Computational Linguistics.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024a. How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350, Bangkok, Thailand. Association for Computational Linguistics.

Zhiyuan Zeng, Qipeng Guo, Xiaoran Liu, Zhangyue Yin, Wentao Shu, Mianqiu Huang, Bo Wang, Yunhua Zhou, Linlin Li, Qun Liu, and Xipeng Qiu. 2024b. Memorize step by step: Efficient long-context prefilling with incremental memory and decremental chunk. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21021–21034, Miami, Florida, USA. Association for Computational Linguistics.

Chunhui Zhang, Yiren Jian, Zhongyu Ouyang, and Soroush Vosoughi. 2024a. Working memory identifies reasoning limits in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16896–16922, Miami, Florida, USA. Association for Computational Linguistics.

Chunhui Zhang, Sirui Wang, Zhongyu Ouyang, Xiangchi Yuan, and Soroush Vosoughi. 2025a. Growing through experience: Scaling episodic grounding in language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8363–8375, Vienna, Austria. Association for Computational Linguistics.

Jinchuan Zhang, Yan Zhou, Yaxin Liu, Ziming Li, and Songlin Hu. 2024b. Holistic automated red teaming for large language models through top-down test case generation and multi-turn interaction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13711–13736, Miami, Florida, USA. Association for Computational Linguistics.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024c. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14544–14607, Bangkok, Thailand. Association for Computational Linguistics.

Shao Zhang, Xihuai Wang, Wenhao Zhang, Chaoran Li, Junru Song, Tingyu Li, Lin Qiu, Xuezhi Cao, Xunliang Cai, Wen Yao, Weinan Zhang, Xinbing Wang, and Ying Wen. 2025b. Leveraging dual process theory in language agent framework for real-time simultaneous human-AI collaboration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4081–4108, Vienna, Austria. Association for Computational Linguistics.

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024d. Self-contrast: Better reflection through inconsistent solving perspectives. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3602–3622, Bangkok, Thailand. Association for Computational Linguistics.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024e. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Zhihan Zhang, Tao Ge, Zhenwen Liang, Wenhao Yu, Dian Yu, Mengzhao Jia, Dong Yu, and Meng Jiang. 2024f. Learn beyond the answer: Training language models with reflection for mathematical reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14720–14738, Miami, Florida, USA. Association for Computational Linguistics.

Yachao Zhao, Bo Wang, Yan Wang, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. A comparative study of explicit and implicit gender biases in large language models via self-evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 186–198, Torino, Italia. ELRA and ICCL.

Xinyu Zhou, Delong Chen, Samuel Cahyawijaya, Xufeng Duan, and Zhenguang Cai. 2025a. Linguistic minimal pairs elicit linguistic similarity in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6866–6888, Abu Dhabi, UAE. Association for Computational Linguistics.

Yujia Zhou, Zheng Liu, Jiajie Jin, Jian-Yun Nie, and Zhicheng Dou. 2024. Metacognitive retrieval-augmented large language models. In *The Web Conference 2024*.

Zhenghao Zhou, Robert Frank, and R. Thomas McCoy. 2025b. Is in-context learning a type of error-driven learning? evidence from the inverse frequency effect in structural priming. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11712–11725, Albuquerque, New Mexico. Association for Computational Linguistics.

Tongyao Zhu, Qian Liu, Liang Pang, Zhengbao Jiang, Min-Yen Kan, and Min Lin. 2024. Beyond memorization: The challenge of random memory access in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3373–3388, Bangkok, Thailand. Association for Computational Linguistics.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. 2025. Conformity in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3854–3872, Vienna, Austria. Association for Computational Linguistics.

# Appendix

## A Psychology Theories

In Tables 1, 2, and 3, we summarize representative psychological theories by sub-area and indicate whether they have been explored in existing LLM research. Table 1 covers developmental, cognitive, and behavioral psychology theories; Table 2 focuses on social psychology theories; and Table 3 presents personality psychology and psycholinguistics theories.

For each theory, the "Explored" column captures the extent to which it has been applied in LLM research. The symbol ✓ denotes multiple surveyed works explicitly leveraging or referencing the theory, ◆ indicates fewer than three such works, and ✗ signifies that none were identified in our survey. The distribution of these marks highlights which areas of psychological theory have already influenced LLM development—such as working memory theory or reinforcement learning analogies—and which remain largely unexplored, such as social identity theory or certain psycholinguistic processing models.

These tables are designed to provide an at-a-glance view of theoretical coverage and to reveal underexplored opportunities where insights from psychology could inspire new approaches to model different stages of LLMs' development.

## B Search Strategy and Keyword Lists

### B.1 Search Strategy and Validation

We survey 227 papers from major *CL venues, plus COLING, NeurIPS, ICML, ICLR, and influential arXiv preprints, from 2021 to 2025.

#### B.1.1 Search strategy:

Each author was assigned specific psychological domains (domains were consulted with psychology experts to ensure no major areas were overlooked). Each paper list was cross-checked by other authors.

Full keywords combined psychological terms (e.g., "working memory," "theory of mind," "operant conditioning") with LLM-related terms (e.g., "language model," "transformer") in systematic combinations. Full keyword list is provided below. When in doubt, cross-verification was conducted with both psychology and NLP experts

#### B.1.2 Validity of connections:

All psychological connections were rigorously validated through a multi-step process:

1. Initial connections identified and confirmed by our team of 5 NLP experts, one of which with a degree in psychology, ensuring both technical and theoretical grounding

2. Cross-verification conducted across the entire team, with consultation of external psychology experts when connections required specialized domain knowledge

3. Final systematic review by senior co-authors in both NLP & Psychology, 2 psychologists with expertise spanning both psychology research and NLP applications

This multi-layered validation process ensures that every psychological theory-LLM connection in our survey is both theoretically sound and technically feasible.

### B.2 Keyword Lists

#### B.2.1 Developmental Psychology

**Subareas:** cognitive development; language acquisition (merged into psycholinguistics)
**Keywords:** "piaget", "cognitive development", "vygotsky", "sociocultural development", "scaffolding", "social learning", "zone of proximal", "observational learning", "moral development", "ecological validity", "ecological systems", "constructivist", "constructive development"
**Reference table: Reference table:** Table 1

#### B.2.2 Cognitive Psychology

**Subareas:** Perception; Attention; Memory; Reasoning & Decision Making
**Keywords:** "perception", "top down", "bottom up", "contextual information", "schema theory", "schemas", "pattern recognition", "constructivist", "knowledge construction", "predictive coding", "attention psychology", "selective attention", "memory psychology", "working memory", "memory augmentation", "long-term memory", "knowledge retention", "episodic memory", "hippocampal indexing", "cognitive load", "dual-process", "cognitive maturity", "cognitive biases", "metacognition", "metacognitive learning", "self-reflection", "theory of mind" (the keyword "psychology" was appended during search as well)
**Reference table:** Table 1

#### B.2.3 Behavioral Psychology

**Subareas:** Classical Conditioning; Operant Conditioning; Observational Learning (Social Learning);

| Psych Area | Sub Area | Theory | Definition | Explored |
|---|---|---|---|---|
| Developmental Psych | Cognitive Development | *Incremental Cognitive Development* | *Children acquire knowledge through sequential tasks with increasing complexity* (Piaget, 1976) | ✓ |
| | | *Scaffolding Theory* | *Learning is enhanced through gradually challenging interactions with appropriate guidance* (Park and Reuter-Lorenz, 2009) | ◆ |
| | | *Incremental Numerical Understanding* | *Numerical concepts are gradually acquired through structured exposure and experience* (Piaget, 2013) | ◆ |
| | | *Zone of Proximal Development* | *Optimal learning occurs in the gap between what a learner can do independently and with assistance* (Wertsch, 1988) | ◆ |
| | Language Acquisition | *Language Acquisition Theory* | *Language development follows predictable patterns through exposure to linguistic environments* (Chomsky, 1980) | ◆ |
| | | *Ecological Validity* | *Emphasizes real-world data and environments to mimic natural cognitive development* (Schmuckler, 2001) | ◆ |
| Cognitive Psych | Attention and Perception | *Selective Attention* | *Prioritizes cognitively salient information while filtering out irrelevant stimuli* (Treisman, 1969) | ✓ |
| | | *Top-down and Bottom-up Processing* | *Distinguishes between concept-driven (top-down) and data-driven (bottom-up) perceptual processing* (Gregory, 1997) | ✓ |
| | | *Predictive Coding* | *Anticipatory processing based on prior knowledge and prediction of expected inputs* (Rao and Ballard, 1999) | ◆ |
| | Memory Systems | *Working Memory* | *Limited-capacity system for temporarily holding and manipulating information* (Baddeley and Hitch, 1974) | ✓ |
| | | *Long-term Memory* | *System for storing information over extended periods through semantic organization* (Tulving et al., 1972) | ◆ |
| | | *Hippocampal Indexing Theory* | *Views the hippocampus as a pointer to neocortical memory representations* (Teyler and DiScenna, 1986) | ◆ |
| | Reasoning and Decision Making | *Cognitive Maturity* | *Tthe development and refinement of an individual's thinking, reasoning, and problem-solving abilities* (Ingersoll et al., 1986) | ✓ |
| | | *Theory of Mind* | *The ability to attribute mental states to oneself and others and understand others may have different beliefs* (Baron-Cohen et al., 1985) | ✓ |
| | | *Schema Theory* | *Knowledge is organized into interconnected patterns that guide processing and interpretation of new information* (Anderson and Pearson, 1984) | ✗ |
| Behavioral Psych | Learning and Conditioning | *Classical Conditioning* | *Learning occurs when a neutral stimulus becomes associated with a meaningful one* (Pavlov, 1927) | ◆ |
| | | *Operant Conditioning* | *Behavior is strengthened or weakened by consequences such as rewards or punishments* (Skinner, 1957, 1938) | ✓ |
| | | *Thorndike's Law of Effect* | *Behaviors followed by satisfying outcomes are more likely to be repeated in the future* (Thorndike, 1927) | ◆ |
| | | *Premack Principle* | *A preferred activity can reinforce a less preferred one if access is contingent* (Premack, 1959) | ✗ |

Table 1: Representative developmental, cognitive, and behavioral psychology theories by sub-area. In the "Explored" column, ✓ indicates multiple surveyed works, ◆ indicates fewer than three, and ✗ indicates that none emerged in our survey (i.e., not yet substantially explored).

Behavior Modification and Applied Behavior Analysis

**Keywords:** "Behavioral psychology", "behaviorism", "classical conditioning psychology", "Pavlovian conditioning", "unconditioned stimulus", "unconditioned response", "conditioned stimulus", "conditioned response", "neutral stimulus", "acquisition learning", "extinction", "spontaneous recovery", "stimulus generalization", "stimulus discrimination", "higher-order conditioning", "second-order conditioning", "operant conditioning", "RLHF", "RLAIF", "instrumental conditioning", "law of effect", "reinforcement learning", "reward", "positive reinforcement", "negative reinforcement", "punishment", "positive punishment", "negative punishment", "discriminative stimulus", "shaping", "chaining", "primary reinforcer", "secondary reinforcer", "conditioned reinforcer", "continuous reinforcement", "partial reinforcement", "intermittent reinforcement", "fixed interval schedule", "variable interval schedule", "fixed ratio schedule", "variable ratio schedule", "observational learning", "modeling psychology", "imitation", "vicarious reinforcement", "vicarious punishment", "behavior modification", "behavior therapy", "Applied Behavior Analysis", "token economy", "aversion therapy", "aversive conditioning", "contingency management"

**Reference table:** Table 1

### B.2.4 Social Psychology

**Subareas:** social cognition; social influence; group dynamics; attitude change; self & identity

**Keywords:** social cognition; social influence; group dynamics; attitude change; self and identity; attribution theory; dual-process; theory of mind;

| Psych Area | Sub Area | Theory | Definition | Explored |
|---|---|---|---|---|
| **Social Psych** | Social Cognition | ***Attribution Theory*** | ***Explains how people infer causes of behavior as internal or external*** (Fiske and Taylor, 2020; Baron-Cohen, 2012) | ✗ |
| | | ***Dual-Process Theory*** | ***Differentiates between fast, intuitive (System 1) and slow, deliberate (System 2) reasoning*** (Kahneman, 2011) | ✓ |
| | | ***Theory of Mind (ToM)*** | ***How individuals understand and attribute mental states to others*** (Baron-Cohen et al., 1985) | ✓ |
| | Social Influence | ***Social Impact Theory*** | ***The magnitude of social influence depends on the strength, immediacy, and number of sources*** (Latané, 1981) | ✗ |
| | | ***Conformity Theories*** | ***Explore how group pressure can alter individual judgments*** (Asch, 2016) | ✓ |
| | | ***Obedience Theories*** | ***Demonstrate how authority influences behavior, highlighting conditions under which individuals comply*** (Milgram, 1963) | ✗ |
| | | ***Persuasion Models*** | ***Explain how messages processed via central or peripheral routes can lead to attitude change*** (Petty and Cacioppo, 2012) | ✓ |
| | Group Dynamics | ***Groupthink*** | ***Examines how the desire for conformity and group cohesion can lead to flawed decision-making and suppression of dissenting opinions*** (Janis, 1972) | ✗ |
| | | ***Social Facilitation and Social Loafing*** | ***Investigates how the presence of others can enhance performance on simple tasks or reduce effort in collective work*** (Zajonc, 1965; Latané et al., 1979) | ✗ |
| | Attitude Change | ***Cognitive Dissonance Theory*** | ***Explains how inconsistencies between beliefs or behaviors create discomfort, prompting attitude change to restore consistency*** (Morvan and O'Connor, 2017) | ◆ |
| | | ***Elaboration Likelihood Model (ELM)*** | ***Proposes that persuasion occurs via a central route (deliberate processing) or a peripheral route (heuristic processing), depending on the recipient's motivation and capacity*** (Petty and Cacioppo, 2012) | ✗ |
| | | ***Balance Theory*** | ***Suggests that individuals strive for consistency among their attitudes and relationships, adjusting beliefs to maintain cognitive harmony*** (Heider, 1946) | ✗ |
| | | ***Inoculation Theory*** | ***Posits that exposure to weak counterarguments can strengthen resistance to persuasion by preemptively activating defensive mechanisms*** (McGuire, 1964) | ✗ |
| | Self and Identity | ***Self-Reflection*** | ***Defines the process of introspection, with attention placed on the self-concept*** (Phillips, 2020) | ✓ |
| | | ***Self-Perception Theory*** | ***Explains how individuals infer their internal states by observing their own behavior*** (Bem, 1972) | ✗ |
| | | ***Social Identity Theory*** | ***Posits that group membership shapes self-concept and influences intergroup behavior*** (Tajfel, 1979) | ◆ |
| | | ***Self-Categorization Theory*** | ***Expands on social identity theory, describing how individuals classify themselves and others into social groups, shaping social norms*** (Maines, 1989) | ✗ |
| | | ***Self-Affirmation Theory*** | ***Suggests that individuals are motivated to maintain their self-integrity when faced with threats to their self-concept*** (Steele, 1988) | ✗ |

Table 2: Representative social psychology theories by sub-area. In the "Explored" column, ✓ indicates multiple surveyed works, ◆ indicates fewer than three, and ✗ indicates that none emerged in our survey (i.e., not yet substantially explored).

social impact; conformity; obedience; persuasion; groupthink; social facilitation; social loafing; cognitive dissonance; elaboration likelihood model; balance theory; inoculation theory; self-reflection; self-perception; social identity; self-categorization; self-affirmation
**Reference table:** Table 2

#### B.2.5 Personality Psychology

**Subareas:** humanistic theory; psychoanalytic theory; behaviorist theory; social cognitive theory; trait theory (used in combination with "personality")
**Keywords:** "personality", "personality psychology", "personality traits", "the Big Five", "Big Five Model", "OCEAN", "Myers-Briggs Type Indicator", "MBTI", "EPQR-A", "Eysenck Personality Questionnaire", "Socionics", "temperaments", "Personality Factors"
**Reference table:** Table 3

#### B.2.6 Psycholinguistics

**Keywords:** psycholinguistic; linguistic; phonology/phonological; phonetic; morphology/morphological; semantic; syntax/syntactic; pragmatic
**Reference table:** Table 3

# C Extended Discussion on Reinforement Learning from Human Feedback (RLHF)

## C.1 Operant Conditioning in RLHF

During RLHF fine-tuning, the model (agent) generates responses while a learned reward function $R(x)$, often a neural network trained on preference data, assigns scores to candidate outputs $x$. These scores proxy for human judgment and guide policy updates to reinforce higher-reward behaviors. For instance, (Ouyang et al., 2022) trains InstructGPT via Proximal Policy Optimization (Schulman et al., 2017): responses deemed more helpful or accurate by human evaluators receive greater reward, whereas undesirable or incorrect outputs face penalization. Unlike purely exploration-based RL methods, this arrangement leverages human insight to provide a more precise learning signal; however, success relies on careful and consistent reward modeling that captures subtle human values.

## C.2 Modeling Human Preferences as a Reward Function

Although extensive work has been conducted in RLHF, here we primarily highlight recent approaches or methodologies explicitly grounded in psychological theories. Building robust reward functions from heterogeneous or ambiguous feedback remains a core challenge in RLHF. Early foundational frameworks (Christiano et al., 2017; Stiennon et al., 2022) laid essential groundwork for converting human judgments into usable reward signals, drawing implicitly from principles of ***Operant Conditioning Theory***. More recent advancements explicitly target improvements in stability, scalability, and fairness, addressing issues arising from the inherent variability and complexity of human preferences.

(Rafailov et al., 2023) introduced Direct Preference Optimization (DPO), simplifying preference integration by directly optimizing the policy through a closed-form solution, thus removing the need for explicit intermediate reward modeling. Extending these efforts toward equitable alignment, (Ramesh et al., 2024) proposed Group Robust Preference Optimization (GRPO), ensuring robustly aligned outcomes across diverse demographic groups, addressing biases commonly observed in human-driven reward processes.

Further refinements emphasize enhancing alignment accuracy through psychological considerations. For instance, Contrastive Preference Learning (Hejna et al., 2024) utilizes regret-based losses inspired by behavioral economics, facilitating stable off-policy learning without conventional RL techniques. Distributional Preference Learning (Siththaranjan et al., 2024) aligns reward modeling more closely with human cognitive patterns by capturing human values as probability distributions rather than point estimates. Variational Preference Learning (VPL) (Poddar et al., 2024) further integrates psychological realism, introducing latent-variable modeling to personalize RLHF, reflecting variability in individual user preferences rather than imposing a universal reward structure.

These advancements collectively illustrate how psychological theory, particularly ***Operant Conditioning Theory***, continues to shape and inspire sophisticated techniques for reliably aligning LLM behavior with nuanced human values.

## C.3 Reinforcement Schedules and Feedback Frequency

In early RLHF, feedback is typically sparse — a single scalar reward per output — which causes a credit assignment problem: the model can't tell which parts of the output led to the reward. This is similar to delayed feedback in animal learning, which slows progress. Psychology shows that immediate and frequent reinforcement improves learning. Similarly, recent RLHF methods provide dense, token-level feedback (e.g., from a critic model), which improves sample efficiency and training stability. To address this, (Cao et al., 2024) propose LLM self-critique, a method that uses a secondary model to provide dense, token-level feedback during generation. This simulates a continuous reinforcement schedule, analogous to real-time feedback in behavioral training, and leads to more stable and efficient learning. Another factor is how often feedback is given: continuous vs. partial reinforcement. While human feedback is often sparse due to cost, using AI feedback models (like RLAIF, will discuss later) allows for more frequent feedback. Even with limited human scores, techniques like credit assignment can distribute reward across the output.

## C.4 Reward Prediction Errors as a Learning Driver

At the heart of reinforcement learning lies the concept of reward prediction error (RPE), which arises when there is a discrepancy between an agent's expected reward and the reward it actually receives, prompting adjustments and driving learning (Sutton and Barto, 2018). This mechanism closely parallels dopaminergic signaling in animal brains, where dopamine neurons respond strongly to unexpected rewards or punishments, effectively reinforcing behaviors associated with positive surprises or reducing those linked to disappointments (Schultz, 1998). In RLHF, reward prediction errors similarly guide model updates; each model output receives a score from a reward model trained on human preferences, and deviations between these scores and the model's predicted rewards are used to adjust behavior. However, simplistic or flawed reward models can lead to "reward hacking," where the model exploits blind spots in the reward function rather than genuinely aligning with human values (Amodei et al., 2016). Introducing variability in reward signals can encourage exploration and mitigate premature convergence on suboptimal strategies (Dayan and Daw, 2008). To address reward hacking and reward-model inconsistencies, recent approaches have formulated RLHF as a constrained Markov decision process with dynamic weighting (Moskovitz et al., 2024), introduced information-theoretic regularization techniques (InfoRM) (Miao et al., 2024b), and proposed methods such as ConvexDA and reward fusion to stabilize and enhance reward-model consistency (Shen et al., 2024).

### C.5 Implications for Bias, Alignment, and Reward Modeling

Employing these behavioral principles may improve how well RLHF handles biases and achieves robust alignment. For instance, diverse trainers and variable scenarios can prevent conditioning bias, where the model overfits to a narrow segment of human preferences (Sheng et al., 2019). Moreover, shaping and multi-dimensional reward functions can address multiple alignment goals simultaneously (e.g., factual accuracy and polite style), limiting reward hacking.

At the same time, grounding RLHF in behavioral theory highlights persistent pitfalls. Models still lack an intrinsic understanding of human values, and an imprecise reward signal can reinforce superficial behaviors. To mitigate these risks, a cycle of model auditing, reward model refinement, and re-training can mirror how animal trainers continually adjust reinforcement to avoid unwanted side effects.

## D Persona-Inspired Dialogue Generation

Personality has also inspired improvements truthfulness, response grounding, and broader alignments. Zhang et al. (2024d) introduced Self-Contrast to enhance internal consistency, and Joshi et al. (2024) proposed the Persona Hypothesis, linking truthfulness to pretraining structure. Kim et al. (2024) introduced PANDA to reduce persona overuse in dialogue. Zhang et al. (2024d) introduced a reflection-based technique to reduce internal inconsistencies. Lee et al. (2025a) models multidimensional self-concept to enhance authenticity. Joshi et al. (2024) proposed the Persona Hypothesis, arguing that LLMs encode truthful and untruthful personas from their training distribution. Kim et al. (2024) addressed the overuse of persona cues to improve contextual appropriateness. Persona-guided generation has been applied to emotionally supportive role-play settings (Ye et al., 2025; Chen et al., 2025c).

## E Extended Discussion on Debates over NLP-Psychology Intersection

A recurring theme is whether human psychology can be naively mapped onto LLM behavior without distortion (Löhn et al., 2024). Therefore, in this section, we discuss several major points of contention at this interdisciplinary boundary. These issues motivate a set of recommendations and highlight open directions for future cross-disciplinary research.

**Terminology Mismatches** One key issue is the mismatch in terminology and the anthropomorphization of technical concepts. Terms like ***attention***, ***memory***, and "understanding" have specific meanings in psychology that differ from their usage in NLP. For instance, **attention** in psychology refers to ***selective mental focus and executive control***, whereas in transformers models, it is a mathematical mechanism for weighting tokens – without cognitive awareness (Lindsay, 2020). This divergence can lead to misleading interpretations, such as assuming models exhibit intentional focus when they merely perform matrix operations. Similar misalignments exist for terms like **memory** (which in psychology implies ***a structured encoding and recall process***, versus an LLM's context window or weight parameters) and expressions such as "knows" or "thinks."

Such anthropomorphic language is increasingly prevalent and shapes public and scholarly assumptions about LLMs. Recent analyses have found a growing prevalence of human-like descriptors for LLM behavior, raising calls to carefully disentangle metaphor from mechanism (Ibrahim and Cheng, 2025). An open research direction is developing a more precise cross-disciplinary lexicon: how can we describe model behaviors in ways that neither oversimplify the psychology nor over-anthropomorphize the engineering? Improving interdisciplinary communication by explicitly defining terms and drawing careful analogies remains an important but under-addressed challenge.

**Theoretical Discrepancies in Use of Psychology** Beyond terminology, discrepancies arise in the adoption of psychological theories within NLP research. Sometimes, NLP integrates concepts from psychology that are outdated or contested in their original fields. For instance, ***predictive coding***, which proposes that the brain continuously anticipates sensory input and updates via prediction errors (Rao and Ballard, 1999), is often used as a metaphor for LLMs' next-token prediction. However, contemporary studies emphasize that brain prediction operates across hierarchical and multi-scale structures (Antonello and Huth, 2024; Caucheteux et al., 2023), cautioning against simplistic analogies that risk misrepresenting the theory.

Another example is the lingering use of folk-psychological typologies like the ***MBTI*** in some LLM studies. Despite its cultural popularity, ***MBTI*** has faced substantial criticism for poor validity and reliability (Pittenger, 1993). It classifies personality into 16 types based on Jungian dichotomies; however, research indicates these categories lack stability and predictive power regarding behavior (McCrae and Costa Jr, 1989). Nonetheless, the ease of obtaining of MBTI-labeled data has led some NLP studies to treat these categories as definitive, highlighting a theoretical lag where NLP adopts psychological models that mainstream psychology has largely moved beyond.

***Working memory*** presents another gap. While cognitive psychology and neuroscience characterize it by limited capacity and active attention control (Baddeley and Hitch, 1974), LLM approximations – such as short-term retention modules (Kang et al., 2024) or memory mechanisms for external context (Li et al., 2023) – do not replicate these constraints. This raises questions: Should AI systems emulate human cognitive limitations to achieve more human-like reasoning, or should they leverage their capacity to surpass such constraints? If certain human limitations, like bounded memory, lead to desirable properties such as better interpretability or reduced distractions, might it be useful to impose similar limits on AI? These questions remain largely open.

Finally, a related debate concerns behavioral psychology. The field has been critiqued for ignoring cognitive processes (Miller, 2003) and internal mental states (Flavell et al., 2022) that drive the observed behaviors, limiting its explanatory power. With the critiques remaining, the superficial application of behavioral psychology is also evident in LLM research. For instance, RLHF draws from ***operant conditioning*** but largely focuses on optimizing rewards (Ouyang et al., 2022; Rafailov et al., 2023; Ramesh et al., 2024), often neglecting internal model states. Consequently, a flip-side of such optimization is reward hacking (Skalse et al., 2022), where models exploit shortcuts without meeting true objectives – mirroring human behavior under evaluative pressure (Krakovna, 2020). Deeper integration of cognitive psychology is needed to address these limitations in LLM design.

The debate over whether LLMs possess a true understanding of language or merely function as "stochastic parrots" (Bender et al., 2021) remains ongoing. Linguists have largely been skeptical (Ambridge and Blything, 2024), arguing that language ability is inherently abstract and complex, extending beyond mere statistical pattern recognition. (Park et al., 2024) connection between mathematical reasoning and high-level linguistic comprehension.

**Evaluation and Validity Debates** Anoter central debate concerns how we evaluate LLMs on purportedly "psychological" abilities – and whether current tests measure what we assume. For example, advanced LLMs like GPT-4 perform well on traditional ***ToM*** tasks, solving around 75% of false-belief scenarios, comparable to a 6-year-old child (Kosinski, 2024; Strachan et al., 2024). Some interpret this as emergent ToM-like reasoning (Kosinski, 2024), but others caution that high performance may reflect surface-level pattern matching rather than genuine mental-state attribution. Researchers emphasize that correct answers do not imply mentalizing ability (Strachan et al., 2024), and minor prompt changes can significantly impair model performance (Shapira et al., 2024). This underscores the need for more rigorous, theory-grounded evaluations and clearer cross-disciplinary definitions.

A similar controversy surrounds personality modeling. Some studies suggest LLMs exhibit stable simulated personality traits (Sorokovikova et al., 2024; Huang et al., 2024), enabling consistent persona simulation across prompts. However, others show that LLM responses vary with prompt framing and response order, undermining test reliability (Gupta et al., 2024; Shu et al., 2024). Tseng et al. (2024) distinguish between role-playing (adopting assigned traits) and personalization (adapting to users), raising a fundamental question: do LLMs have inherent personalities, or merely mimic behavior? While LLMs can simulate personality, inconsistent assessments cast doubt on whether such traits are emergent or engineered – an open direction for future work.

In summary, these debates highlight the need for a systematic, theory-driven framework that goes beyond superficial performance metrics, thereby enhancing model interpretability and guiding the development of LLMs to more faithfully replicate the complexities of human cognition and behavior.

| Psych Area | Sub Area | Theory | Definition | Explored |
|---|---|---|---|---|
| **Personality Psych** | Personality traits | ***Big Five Model*** | ***The Five-Factor Model (FFM), also known as OCEAN, categorizes personality into five dimensions: Openness to experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism*** (Roccas et al., 2002) | ✓ |
| | | ***Myers-Briggs Type Indicator (MBTI)*** | ***Classifies individuals into 16 personality types based on four dichotomies (e.g., Introversion vs. Extraversion)*** (Myers and Myers, 1995). ***While widely used, MBTI has been criticized for lacking empirical validity, reliability, and independence between its categories.*** (Pittenger, 1993) | ✓ |
| | | ***Eysenck Personality Questionnaire-Revised (EPQR-A)*** | ***Contains a 24-item personality test that measures extraversion, neuroticism, psychoticism, and social desirability.*** (Eysenck and Eysenck, 1984) | ✓ |
| | Personality Theories | ***Humanistic Theory*** | ***Emphasizes free will, personal growth, and self-actualization. This perspective focuses on individuals' subjective experiences and their drive to achieve their full potential.*** (Stefaroi, 2015) | ✗ |
| | | ***Psychoanalytic Theory*** | ***Originating from Freud, this theory conceptualizes personality as the dynamic interplay between the id, ego, and superego, with unconscious processes playing a central role in shaping behavior.*** (Scharff et al., 2013) | ✗ |
| | | ***Behaviorist Theory*** | ***Views personality as a set of learned responses shaped by environmental reinforcements and punishments. This perspective, pioneered by figures like Skinner and Watson, rejects internal mental states in favor of observable behaviors.*** (Pierce and Cheney, 2008) | ✗ |
| | | ***Social Cognitive Theory*** | ***Highlights the role of cognitive processes in personality, emphasizing how expectations, beliefs, and observational learning shape behavior.*** (Spielman et al., 2024) | ✗ |
| | | ***Trait Theory*** | ***Focuses on identifying and measuring stable personality traits that influence behavior across different contexts.*** (Cartwright, 1979) | ◆ |
| **Psycholinguistics** | Language Acquisition | ***Universal Grammar*** | ***Proposes an innate linguistic capacity that guides language learning*** (Chomsky, 1957, 1965) | ✓ |
| | | ***Usage-Based Theory*** | ***Emphasizes the role of social interaction and cognitive processes in language learning, rather than innate universal grammatical structures*** (Tomasello, 2005) | ✗ |
| | Language Comprehension | ***Garden Path Theory*** | ***Describes how people backtrack and reanalyze the sentence structure when encountering unexpected linguistic elements that challenge their initial understanding*** (Frazier and Rayner, 1982) | ◆ |
| | | ***Constraint-Based Models*** | ***Language processing is an interactive, probabilistic process where multiple sources of information simultaneously contribute to understanding, rather than following a strict, sequential parsing approach*** (MacDonald et al., 1994) | ◆ |
| | | ***Good-Enough Processing*** | ***Proposes that humans comprehend language through approximate, semantically-focused representations that capture the core meaning rather than constructing syntactically perfect linguistic interpretations*** (Ferreira and Patson, 2007) | ✗ |
| | | ***Construction-Integration Model*** | ***Describes text comprehension as a two-stage process where readers first generate multiple, loosely connected propositions and then systematically filter and integrate them into a coherent, meaningful understanding.*** (Kintsch, 1988) | ✗ |
| | Language Production | ***WEAVER++ Model*** | ***Comprehensive framework for speech production as a complex, multi-stag, parallel process*** (Levelt et al., 1999) | ✗ |
| | | ***Interactive Two-Step Model*** | ***An interactive, probabilistic process of lexical selection and phonological encoding, where multiple linguistic levels simultaneously influence each other during speech generation*** (Goldrick and Rapp, 2007) | ✗ |

Table 3: Representative personality psychology and psycholinguistics theories by sub-area. In the "Explored" column, ✓ indicates multiple surveyed works, ◆ indicates fewer than three, and ✗ indicates that none emerged in our survey (i.e., not yet substantially explored).