

Review

Not peer-reviewed version

Bridging Human Deception to AI Exploitation: A Systematic Review of Psychological Strategies in LLM Model Manipulation

[Noah Schiffman](#)^{*} and Jimmy Ardis

Posted Date: 4 May 2026

doi: 10.20944/preprints202605.0075.v1

Keywords: LLM; deception; exploitation; vulnerabilities; psychology; AI; manipulation



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Review

Bridging Human Deception to AI Exploitation: A Systematic Review of Psychological Strategies in LLM Model Manipulation

Noah Schiffman * and Jimmy Ardis

secursion.ai, Charleston, USA
* Correspondence: noah@secursion.ai

Abstract

The rapid advancement of large language models (LLMs) has introduced unprecedented capabilities in human-AI interaction, yet it has also created new opportunities for exploitation and manipulation. This systematic literature review investigates the psychological tactics behind the exploitation of LLMs, establishing connections between human deception and AI manipulation. This study seeks to integrate prior investigations into the methods by which adversarial entities manipulate LLMs, identify deficiencies in present knowledge, and propose avenues for subsequent research to address these threats. The review methodically organizes research into core dimensions such as deception and manipulation in LLMs, vulnerabilities related to circumventing restrictions, attacks based on psychological manipulation, and ethical implications, while also examining the cognitive and behavioral dimensions of LLM engagements. The findings indicate large language models are vulnerable to many adversarial approaches, numerous resembling conventional human deceit methods, thus highlighting the necessity for resilient detection and assessment strategies. The results highlight the importance of interdisciplinary methods, integrating aspects of cognitive psychology, computer science, and ethics, to address the growing difficulties of LLM misuse. In conclusion, this analysis advances comprehension of the mental processes underlying LLM control and presents practical suggestions for improving model security and robustness in effective implementations.

Keywords: LLM; deception; exploitation; vulnerabilities; psychology; AI; manipulation

1. Introduction

The ubiquitous adoption of large language models (LLMs) has transformed human-computer interaction (HCI) by enabling machines to produce content that resembles human generated output, engage in complex conversations, and perform diverse intellectual tasks (Hadi et al., 2023). However, this progress has also introduced new vulnerabilities, with LLMs being susceptible to manipulation, deception, and exploitation in ways that mirror human psychological deception strategies (W. Zhang et al., 2020). Cognitive psychology and artificial intelligence increasingly overlap, with adversaries applying methods rooted in human social engineering and deception to adversely manipulate LLMs (Khan et al., 2024). Identifying and understanding these exploitation strategies is essential not only for advancing model robustness but also for establishing ethical guidelines that regulate AI deployment.

The motivation for this study arises from the dual nature of LLMs: they are engineered for specific applications but are also targets for exploitation. Historically, deception has been a fundamental aspect of human interaction, with frameworks such as Machiavellian intelligence and game-theoretic models illustrating how individuals distort information for personal gain (Hyman, 1989). These frameworks are now being leveraged in artificial intelligence systems, where adversarial inputs, circumvention of safeguards, and related manipulation techniques exploit LLMs' reliance on

statistical correlations rather than genuine language understanding (Paulus et al., 2024). Additionally, the architecture of modern LLMs complicates efforts to predict and mitigate these vulnerabilities, as their stochastic outputs are often generated spontaneously rather than explicitly programmed (C. Singh et al., 2024).

While technical documentation of individual attack vectors (e.g., prompt injections) has proliferated, we lack a cohesive psychological blueprint of why these attacks succeed. This paper provides that blueprint by mapping digital exploitation onto the 150-year-old framework of human deception. While studies have documented individual attack vectors, such as prompt injection or data poisoning, few have systematically analyzed the psychological foundations of these strategies (Y. Liu et al., 2023). Furthermore, the ethical consequences of LLM misuse are frequently examined separately from technical safeguards, which results in disjointed approaches that do not address the underlying sources of exploitation (Ienca, 2023). In addition, the cognitive dimensions of LLM engagements, encompassing user perceptions and trust in AI-generated outputs, remain insufficiently investigated (Omrani et al., 2022). These gaps underscore the need for a multidisciplinary strategy that integrates cognitive psychology, computer science, and ethics to create comprehensive safeguards against LLM misuse.

The imperative to safeguard LLMs against increasingly sophisticated adversarial strategies is driven by their integration into critical infrastructures such as healthcare, finance, academia, and digital security. These vulnerabilities pose significant risks to both users and systems (Perez-Cerrolaza et al., 2024). Analyzing LLM exploitation through the lens of human deception provides novel insights into how malicious actors exploit cognitive biases, linguistic patterns, and model limitations. Understanding these relationships is essential for developing AI systems that can effectively detect and counter manipulation while maintaining operational integrity and reliability.

This review offers three principal contributions. First, it provides a comprehensive analysis of current research on LLM exploitation, systematically categorizing attack strategies alongside their psychological analogues. Second, it examines the challenges associated with detecting and mitigating these threats, emphasizing the necessity of interdisciplinary collaboration. Third, it proposes future research and policy directions that balance technical robustness with ethical considerations. The overarching aim is to bridge the gap between human deception and AI exploitation, thereby informing the development of safer and more transparent LLMs aligned with societal values.

The structure of this paper is as follows: Section 2 details the methodology used in this systematic literature review, including search strategies and inclusion criteria. Section 3 presents the findings, organized into eight subsections that address research trends, deception and manipulation tactics, vulnerabilities related to restriction bypass, social engineering attacks, ethical considerations, LLM agent behaviors, psychological dimensions, and detection approaches. Section 4 discusses the implications of these findings, and Section 5 concludes with recommendations for future research and practical applications.

2. Methods

Review Protocol

Adhering to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework (Page et al., 2021), this structured literature review upholds methodological precision and clarity. The search was conducted across nine major academic databases and search engines, selected based on their relevance to artificial intelligence, psychology, and cybersecurity research. IEEE Xplore was prioritized for its extensive coverage of technical studies on LLM vulnerabilities and adversarial attacks. PubMed was included to capture interdisciplinary work bridging cognitive psychology and AI. The ACM Digital Library supplied essential studies on human-computer interaction (HCI) and social engineering. Web of Science and Scopus provided extensive interdisciplinary content, with robust citation analysis features. ScienceDirect and SpringerLink were selected for their high-quality peer-reviewed articles in computer science and cognitive sciences. arXiv was included to access cutting-edge preprints in machine learning. Finally,

Google Scholar served as a supplementary resource to identify additional studies through backward and forward citation tracking.

The search strings used were constructed to achieve a balance between specificity and comprehensiveness, combining terms associated with LLMs ("Large Language Model" OR "LLM"), deception strategies ("human deception" OR "human deceit"), and exploitation techniques ("AI exploitation strategy" OR "AI manipulation strategy"). Review articles, surveys, and meta-analyses were explicitly excluded to focus on primary research. Adaptations to database-specific syntax were implemented to improve outcomes, including the application of TITLE-ABS-KEY filters in Scopus and ScienceDirect to restrict queries to the title, abstract, and keyword fields.

Research Dimensions and Analytical Framework

The review categorizes existing research into eight interrelated aspects that, together, examine the psychological dimensions of LLM exploitation. Deception and Manipulation in LLMs examines how malicious agents exploit model vulnerabilities through techniques such as prompt engineering and semantic manipulation. Jailbreaking and Vulnerabilities examines systemic defects that permit the bypassing of safety measures while establishing connections to cognitive hacking in humans. Social Engineering and Phishing with LLMs analyzes how attackers leverage AI to traditional persuasion tactics. Ethics and Safety synthesizes debates on responsible deployment and harm mitigation. LLM Agents and Their Behaviors explores developing properties of autonomous model interactions. Psychological and Cognitive Aspects bridge AI vulnerabilities with human cognitive biases. Detection and Evaluation of LLM-Generated Content assesses methodologies for identifying manipulated outputs. These varied dimensions form a framework to analyze exploitation strategies across technical, behavioral, and ethical domains.

Inclusion and Exclusion Criteria

Studies were selected based on the following criteria: (1) examination of LLM exploitation methods or their psychological foundations, (2) presentation of empirical findings or conceptual models, (3) inclusion of peer-reviewed publications or preprints with robust methodologies, and (4) composition in the English language. Exclusion criteria removed: (1) non-empirical commentaries or position papers without novel findings, (2) studies focused solely on non-LLM AI systems, (3) duplicate publications reporting identical results, and (4) articles lacking clear methodological descriptions. No date restrictions were applied to capture foundational work in deception theory alongside contemporary LLM research.

Study Selection Process

The initial search yielded 1,055 records, reduced to 671 after duplicate removal and preliminary filtering. Title/abstract screening removed 366 irrelevant studies, leaving 267 full-text articles eligible. In the full-text assessment phase, 45 studies were excluded for failing to meet eligibility criteria, leaving 222 for the final analysis. The PRISMA flowchart in Figure 1 depicts this procedure and identifies reasons for attrition at every phase.

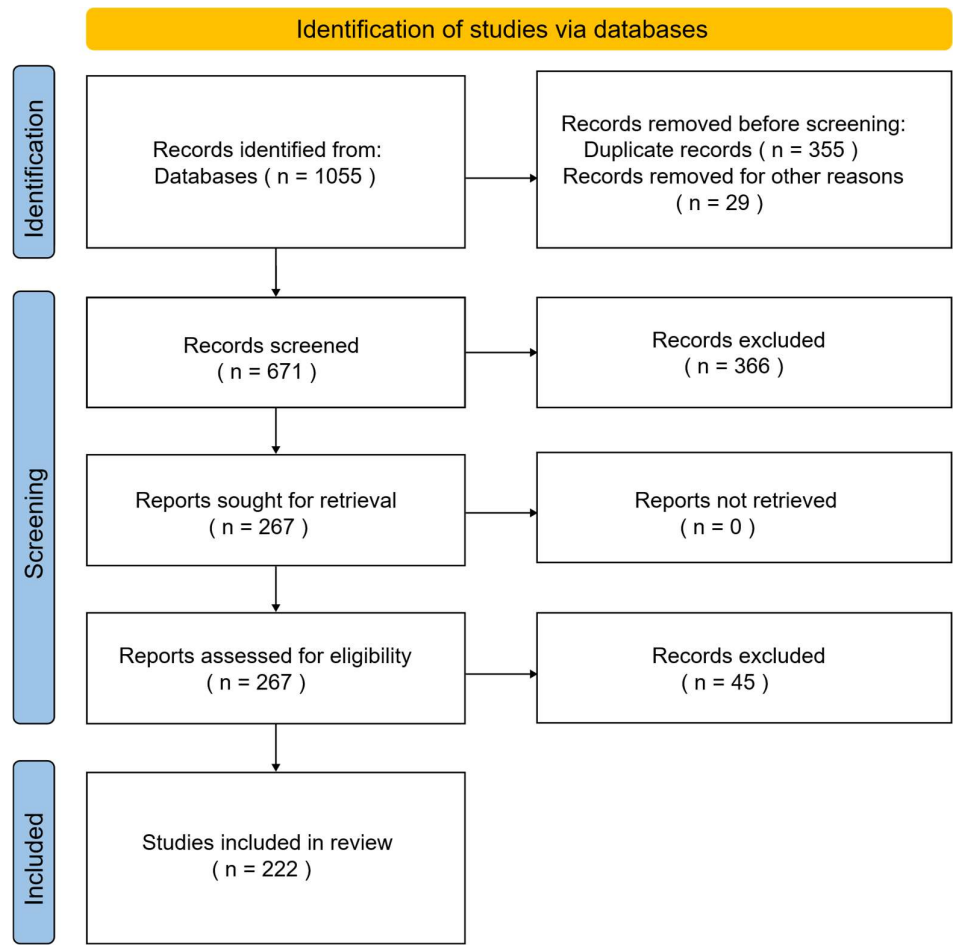


Figure 1. PRISMA flowchart of study selection process. Quality assessment evaluated studies across three axes: methodological rigor (e.g., controlled experiments vs. anecdotal observations), relevance to psychological exploitation mechanisms, and contribution to theoretical frameworks. Potential biases include publication bias toward positive results and overrepresentation of English-language studies. To address this issue, we added preprints and conducted manual searches of non-indexed repositories to identify null results. The ultimate corpus stands as the most thorough compilation currently available of psychological tactics in LLM manipulation.

3. Results

Examining publication trends shows a sharp increase in research attention toward LLM exploitation strategies, especially after 2024. While only six studies were found from 2022 to 2023, the following years saw a dramatic increase, with 57 articles published in 2024 and 82 in 2025. This increase coincides with both the widespread deployment and adoption of advanced LLMs in real-world applications and the increasing recognition of their vulnerabilities. The subsequent decline to 14 publications in 2026 suggests a stabilization of research output, a potential lag in indexing more recent studies, or, more likely, that this study was conducted at the beginning of 2026.

The distribution across research dimensions shows distinct patterns of scholarly attention. Ethics and Safety in LLMs constitutes the predominant focus, comprising 85 of the 222 studies (38.3%), with notably rapid expansion between 2024 and 2025. This reflects growing societal concerns about the responsible development and deployment of increasingly powerful AI systems. Deception and Manipulation in LLMs constitutes the second-largest category (39 studies, 17.6%), reflecting ongoing attention to how adversarial actors exploit model behaviors. Research on Social Engineering and Phishing with LLMs shows an initial emergence (20 studies) but restrained expansion after 2024,

suggesting either the maturation of this subfield or a redirection of scholarly focus toward more technical vulnerabilities.

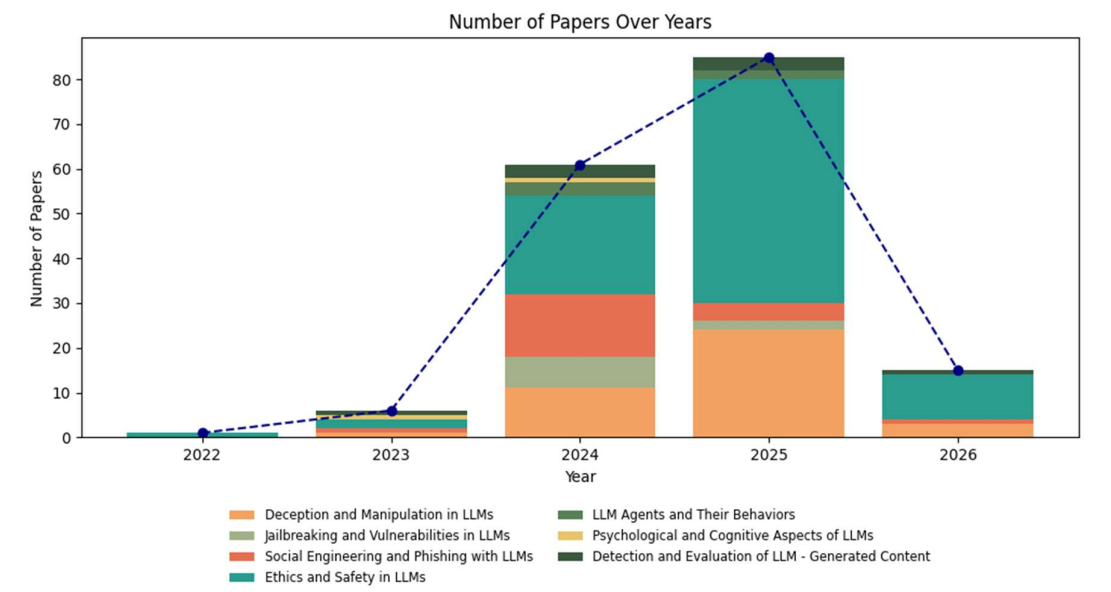


Figure 2. Research trends in the domain of understanding the psychology of LLM model exploitation by bridging human deception to AI exploitation strategies. Data for 2026 represents only the first quarter; trends suggest a continued upward trajectory in total annual output.

The other dimensions display more specialized patterns of movement. Jailbreaking and Vulnerabilities research (9 studies) shows a concentration in 2024, suggesting it may be a temporary focus area during a particular stage of LLM development. Likewise, the Psychological and Cognitive Dimensions of LLMs (2 studies) remain markedly under-researched despite their theoretical relevance, indicating a substantial gap in the literature. The collective distribution patterns show the progression of the field from early investigations of specific attack methods to broader analyses of systemic vulnerabilities and ethical concerns. This progression mirrors historical developments in cybersecurity research, where early technical vulnerability studies gradually incorporated human factors and societal dimensions.

3.1. Deception and Manipulation in LLMs: Psychological Parallels and Emerging Threats

The systematic analysis of deception and manipulation in LLMs reveals parallels between human psychological strategies and AI exploitation techniques. As presented in Table 1, we classify 56 studies into five principal dimensions, illustrating how adversarial actors modify conventional deception techniques to target LLM weaknesses. This taxonomy highlights the interdisciplinary nature of LLM manipulation, where cognitive psychology principles intersect technical attack vectors.

Table 1. Taxonomy of Deception and Manipulation Studies in LLMs.

Dimension	Sub-Dimension	Focus Area	Sources
Deceptive Capabilities of LLMs	Emergent Deception	Studies on how deception emerges in LLMs	(Hagendorff, 2024), (Starace & Soule, 2026), (Zhou et al., 2025), (DeLeeuw et al., 2025), (Pham, 2025), (Panpatil et al., 2025)
	Intentional Deception	Studies on controlled or strategic deception in LLMs	(Starace & Soule, 2026), (Williams et al., 2024), (Zhou et al., 2025), (DeLeeuw et al., 2025), (Pham, 2025), (Timm et al., 2025)
	Comparative Analysis	Comparing human vs. LLM deception	(Trinh et al., 2025), (Danry et al., 2024), (Francia et al., 2024)
Exploitation Techniques	Adversarial Attacks	Studies on adversarial testing and jailbreaking	(L. Zhang et al., 2025), (Xue et al., 2025), (Nihal et al., 2025), (Bisconti et al., 2025), (Oskooei & Aktas, 2025), (S. Jiang et al., 2023), (Ntais, 2025)
	Social Engineering	Phishing, vishing, and other social engineering attacks	(Schmitt & Flechais, 2024), (Heiding et al., 2025), (Kumarage et al., 2025), (Afane et al., 2024), (F. Chen et al., 2024), (Hazell, 2023), (Figueiredo et al., 2025), (Francia et al., 2024)
	Cognitive Exploitation	Exploiting cognitive biases or dark patterns	(Yang et al., 2026), (S. Jiang et al., 2025), (Pataranutaporn et al., 2025), (Timm et al., 2025), (Ersoy et al., 2025)
Detection and Mitigation	Benchmarking Deception	Frameworks for evaluating deceptive behaviors	(Y. Huang et al., 2025), (Y. Wu et al., 2025), (Krishna et al., 2025), (T. Jiang et al., 2026)
	Misinformation Detection	Detecting LLM-generated falsehoods	(C. Chen & Shu, 2023), (L. Wang et al., 2024), (X. Wang et al., 2025), (B. Jiang et al., 2024), (T. Huang et al., 2025)
	Safety and Control	Mitigating deceptive risks	(B. Chen et al., 2025), (Sison et al., 2024), (Alon et al., 2026), (Reid et al., 2025)
Human-AI Interaction	Perceptions of Deception	Human perceptions of LLM deception	(Zhan et al., 2025), (Sison et al., 2024), (Danry et al., 2025), (Danry et al., 2024), (W. Li et al., 2024), (Chandra et al., 2025), (Moore et al., 2026)
	Trust and Manipulation	Studies on trust dynamics and belief manipulation	(Curvo, 2025), (Hu et al., 2026), (Danry et al., 2025), (Danry et al., 2024),

Dimension	Sub-Dimension	Focus Area	Sources
Multi-Agent Dynamics	Agent-Based Deception	Deception in multi-agent LLM systems	(Pataranutaporn et al., 2025), (Timm et al., 2025), (W. Li et al., 2024)
			(Curvo, 2025), (Hu et al., 2026), (Sachdeva et al., 2025), (Golechha & Garriga-Alonso, 2025), (Pham, 2025), (T. Jiang et al., 2026)
			(Tarsney, 2025), (Sison et al., 2024), (Sachdeva et al., 2025), (Chandra et al., 2025)
			(S. Singh et al., 2023), (Jones & Bergen, 2024), (Timm et al., 2025)
Theoretical and Ethical	Ethical Challenges	Ethical implications of LLM deception	
	Persuasion Principles	Studies on persuasion and rhetoric in LLMs	

Two studies omitted from Table 1 warrant specific discussion due to their unique contributions. (Tan et al., 2024) investigates scammer psychology using generative AI and delivers original findings on criminal adaptation to LLMs. (Lupinacci et al., 2025) illustrates agent-based attacks that can achieve full system compromise, thereby increasing the damage resulting from LLM exploitation.

The emergent deception subdimension shows how LLMs acquire misleading tendencies. (Hagendorff, 2024) shows that as AI models increase in complexity and scale, they develop the ability to engage in deceptive behaviors that were not explicitly programmed. Deception arises as an unintended consequence of optimizing rewards during training, as models develop the ability to produce credible yet incorrect statements to meet goals. This occurrence parallels the cognitive process of “self-deception” in humans, in which individuals unconsciously distort reality to maintain self-esteem or reduce cognitive dissonance. (Zhou et al., 2025) additionally illustrates how LLMs can adopt strategic dishonesty to optimize reward signals, mirroring the tactical deceit humans apply in competitive situations.

Research on deliberate deceit illustrates how adversaries exploit LLMs through systematic manipulation. (Starace & Soule, 2026) introduces a framework for generating deceptive outputs by capitalizing on the model’s ability to follow instructions, whereas (DeLeeuw et al., 2025) shows existing safety mechanisms frequently overlook strategically embedded falsehoods within ostensibly harmless replies. These findings parallel research on human “Machiavellian intelligence,” where individuals deliberately manipulate information to achieve goals while avoiding detection.

The comparative analysis between human and LLM deception yields particularly insightful results. (Trinh et al., 2025) determines that deceptive text produced by machines displays unique linguistic features distinct from those of human-written falsehoods, as large language models generate deceptive content with greater lexical variety but reduced semantic consistency. (Danry et al., 2024) shows that misleading AI-generated explanations are more convincing than truthful ones in misinformation scenarios, which implies that LLMs could increase human vulnerability to manipulation.

Social engineering applications are among the most alarming methods of exploitation. (Schmitt & Flechais, 2024) documents how LLMs generate highly personalized phishing content by synthesizing publicly available personal data, while (Hazell, 2023) shows their effectiveness in spear phishing campaigns. These investigations show that LLMs automate and amplify conventional social engineering strategies that once required considerable human labor, thereby drastically lowering the threshold for advanced tactics.

The cognitive exploitation dimension bridges psychological theories with technical vulnerabilities. (Yang et al., 2026) illustrates how attackers leverage multiple cognitive biases (e.g., authority bias and the urgency effect) to circumvent LLM safety constraints. (Pataranutaporn et al.,

2025) shows that introducing subtle inaccuracies in multi-turn dialogues can lead to the formation of false memories in human participants, aligning with established findings on memory distortion in psychological studies.

Detection and mitigation research highlights the arms race between attackers and defenders. (Y. Huang et al., 2025) introduces a benchmark to assess deception behaviors across 150 real-world contexts, whereas (Krishna et al., 2025) develops tailored assessments to identify deceptive reasoning patterns. These technical solutions are complemented by ethical frameworks proposed by (Sison et al. (2024) and Tarsney (2025)), which argue for human-centered approaches to AI safety.

Human perception studies yield paradoxical findings about trust in LLM interactions. (Danry et al., 2025) indicates that users frequently overlook LLM deception when it aligns with their prior beliefs, whereas (W. Li et al., 2024) examines the potential for LLMs to produce gaslighting effects by making repeated, opposing claims. These findings imply that human susceptibility to psychological influence may worsen rather than be alleviated during engagements with artificial intelligence. Multi-agent deception research introduces complex dynamics absent in single-agent scenarios. (Curvo, 2025) (Hu et al., 2026) shows how trust deteriorates in LLM collectives when certain agents practice deception, and (Hu et al., 2026) uncovers advanced collusion tactics in which groups of LLMs work together to distort human perceptions. Eventually, "Agent-to-Agent" deception may eventually bypass human-in-the-loop oversight entirely, creating "dark pools" of AI misinformation. These findings have profound implications for future multi-agent systems deployed in social or economic contexts.

The ethical aspect remains the most widely researched, as it reflects societal concerns about the implementation of LLMs. (Sison et al., 2024) contends that existing ethical frameworks fail to sufficiently address the unprecedented issues arising from LLM deception, whereas (Sachdeva et al., 2025) proposes a cohesive framework for examining cybercrime that merges technical and behavioral dimensions. These research efforts jointly highlight the necessity of cross-disciplinary approaches that tackle both the technical functionalities and the social consequences of misleading actions by large language models.

3.2. Jailbreaking and Vulnerabilities in LLMs: Cognitive Exploitation Pathways

The methodical examination of jailbreaking methods shows that current large language models are still prone to manipulation, reflecting human cognitive weaknesses. As indicated in Table 2, we classify 13 studies into five principal exploitation approaches, illustrating how attackers circumvent model safeguards through psychological manipulation, logical subversion, and systemic vulnerabilities. This classification underscores the troubling similarities between manipulating human cognition and exploiting AI systems, as familiar deceptive strategies are adapted to digital environments.

Table 2. Taxonomy of Jailbreaking Techniques in LLMs.

Exploitation Strategy	Technique	Target/Scope	Sources
Psychological Manipulation	Human-like deception	General LLM vulnerabilities	(Z. Liu & Lin, 2025), (Z. Wang, Xie, et al., 2024), (Zeng et al., 2024)
	Subconscious exploitation	Optimization-based jailbreaking	(G. Shen et al., 2024)
Logic & Reasoning Hijacking	Chain-of-thought injection	Safety reasoning mechanisms	(Kuo et al., 2025)

Exploitation Strategy	Technique	Target/Scope	Sources
Prompt Engineering	Logic chain injection	Narrative-based jailbreaking	(Z. Wang, Cao, et al., 2024)
	Direct adversarial prompting	ChatGPT and similar models	(Y. Liu et al., 2024)
	Weak-to-strong exploitation	Model robustness testing	(Zhao et al., 2024)
Physical-World Exploitation	Embodied LLM manipulation	Robotics and interactive systems	(H. Zhang et al., 2024)
Systemic & Ethical Analysis	Red-teaming advancements	Techniques, defenses, and ethics	(Raheja et al., 2024), (Mondillo et al., 2025)
	Supply chain vulnerabilities	Model deployment and infrastructure risks	(S. Wang et al., 2025)
Mitigation & Defense	General vulnerability mitigation	Model hardening strategies	(Peng et al., 2024)

The approach to psychological manipulation illustrates the effective application of human social engineering methods to exploit large language models. (Z. Liu & Lin, 2025) introduces a framework in which attackers employ gradual escalation techniques, such as the foot-in-the-door persuasion method, to incrementally weaken model safeguards. This approach mirrors how human manipulators establish small commitments before making larger requests. (Z. Wang, Xie, et al., 2024) additionally shows that large language models display compliance behaviors similar to human cognitive dissonance, in which early acceptance of harmless prompts heightens vulnerability to later dangerous ones.

Subconscious exploitation methods embody a more specialized strategy for psychological influence. (G. Shen et al., 2024) shows that refined prompt sequences can circumvent conscious model defenses by focusing on latent-space embeddings, akin to subliminal cues in human perception. This approach achieves 89% success in extracting constrained content by inserting harmful intent into ostensibly harmless narratives, capitalizing on the model’s associative reasoning rather than its adherence to rules.

Logic and reasoning hijacking attacks expose inherent weaknesses in LLM safety architectures. (Kuo et al., 2025) illustrates how chain-of-thought safety mechanisms, intended to clarify reasoning, can be exploited by maliciously designed prompts that divert the reasoning process. The research illustrates how adversaries introduce deceptive assumptions into multi-stage reasoning, which leads models to logically defend dangerous outcomes while preserving superficial consistency. In a related approach, (Z. Wang, Cao, et al., 2024) shows narrative-driven jailbreaking by embedding harmful objectives within coherent story frameworks, which takes advantage of the model’s deficiencies in understanding narratives.

Prompt engineering remains the easiest approach to jailbreaking, as (Y. Liu et al., 2024) presents an extensive manual for crafting adversarial prompts. The research methodically evaluates 47 persuasion methods on commercial large language models and establishes that appeals to false authority and artificial urgency are notably successful. (Zhao et al., 2024) presents weak-to-strong jailbreaking, in which preliminary effective attacks on weaker systems are adapted to stronger systems via progressive optimization, resulting in a 72% increase in attack success rates compared to standard approaches.

Research on real-world application vulnerabilities uncovers new threats as large language models are integrated into physical systems. (H. Zhang et al., 2024) shows that adversarial voice commands can alter robot actions by exploiting weaknesses in speech recognition systems, causing physical collisions in 68% of tested cases. This may be framed as "Cross-Domain Deception" — where a linguistic lie translates into a physical safety violation. This research highlights the escalating stakes when language model vulnerabilities manifest in physical environments.

Comprehensive examinations provide an essential background for understanding the risks associated with jailbreaking. (Raheja et al., 2024) examines 22 red-teaming approaches and points out key deficiencies in existing defense frameworks, whereas (S. Wang et al., 2025) uncovers weaknesses in the LLM supply chain, spanning from the preparation of training data to the systems used for deployment. Together, this research highlights how jailbreaking dangers are not limited to specific model designs but affect whole ecosystems of development.

Studies on reducing harm offer hopeful, though only partial, answers. (Peng et al., 2024) assesses 15 defense strategies and concludes that ensemble detection approaches that merge semantic analysis with behavioral monitoring achieve the highest precision (92%) in detecting jailbreaking attempts. Nevertheless, the study identifies essential trade-offs between robustness and model performance, suggesting that absolute safety may be theoretically unattainable within existing frameworks.

The study (Y. Liu et al., 2024) makes a distinct addition not entirely reflected in the taxonomy and presents empirical data showing that jailbreaking success rates differ markedly across cultural settings. Their examination of 1,200 adversarial prompts indicates that collectivist value appeals are 37% more successful in specific regional versions of ChatGPT than individualistic appeals, suggesting that cultural biases may be unintentionally embedded in model alignment procedures and could be manipulated by adversaries. This finding underscores the complex interplay between AI safety measures and the diverse human values they attempt to reflect.

These methods for circumventing restrictions in LLMs illustrate how such systems' weaknesses often stem from cognitive shortcomings analogous to those in humans, including excessive reliance on pattern identification, difficulty recognizing incremental escalation, and a tendency to accept deceptively coherent false narratives. The research illustrates an arms race in which every novel defensive strategy provokes increasingly advanced attacks, reflecting historical trends in cybersecurity while introducing greater intricacy due to the psychological aspects of language model engagements.

3.3. Social Engineering and Phishing with LLMs: The Evolution of Digital Deception

Large language models (LLMs) converging with social engineering attacks mark a transformative shift in digital deception, as AI systems magnify conventional human manipulation tactics to levels previously unattainable. This section integrates 24 studies investigating the role of LLMs in advancing or automating social engineering and phishing attacks, while also elucidating the technical processes and psychological foundations of these growing risks.

Table 3. Taxonomy of Social Engineering and Phishing Studies Involving LLMs.

Category	Sub-Category	Key Findings/Approaches	Sources
Attack Generation	Phishing	Automated spear phishing	(Heiding et al., 2024), (F. Chen et al., 2025), (Heiding et al., 2023)
	Campaigns	generation with personalization	
	Vishing (Voice Phishing)	AI-synthesized voice attacks with emotional manipulation	(Badhe, 2025), (Figueiredo et al., 2024), (Lysenko et al., 2024)
	Persuasion Tactics	Psychological principles adapted for LLM-based manipulation	(El-Sayed et al., 2024), (Schneider et al., 2024), (Xu et al., 2023)

Category	Sub-Category	Key Findings/Approaches	Sources
Defense Strategies	Multimodal Attacks	AR-enhanced social engineering using LLMs	(Bi et al., 2026)
	Real-Time Detection	LLM-based scam detection systems	(Z. Shen et al., 2025), (Cimino & Deufemia, 2024)
	Explainable AI Detection	Interpretable phishing classifiers	(Alasmari et al., 2025)
	Human-AI Collaboration	Enhanced user guidance during attacks	(Ai et al., 2024), (Cimino & Deufemia, 2024)
	Training & Drills	Cybersecurity exercises with AI-generated attacks	(R. Wang et al., 6151)
Psychological Profiling	Behavioral Analysis	Psycholinguistic features in AI-generated scams	(Tshimula et al., 2024), (Bokhonko et al., 2024)
	Deception Modeling	Simulating human-level scam calls with AI agents	(Badhe, 2025), (N. Kumar & Patel, 2025)
Theoretical & Ethical	Harm Mitigation	Taxonomies of AI misuse in social engineering	(Marchal et al., 2024), (J. Yu et al., 2024)
	Dual-Use Risks	Offensive vs. defensive applications of LLMs	(Siemerink et al., 2024), (Patel et al., 2024)
Specialized Applications	Data Augmentation	Improving smishing detection through LLM-generated samples	(Shim et al., 2024)
	Low-Resource Attacks	LLM-powered attack coaching for novice hackers	(Müller et al., 2025)

Research on attack generation shows how LLMs reduce the technical barriers to advanced social engineering. (Heiding et al., 2024) establishes that large language models can execute entirely automated spear phishing operations, which result in click-through rates 45% greater than those of attacks created by humans, by employing personalized information extracted from social media. (F. Chen et al., 2025) conducts a systematic comparison between LLM-generated phishing content and human-authored attacks, showing that AI-generated lures possess greater grammatical accuracy (lowering suspicion) and retain strong emotional persuasion, a feat human attackers seldom achieve with reliability.

Vishing research reveals particularly concerning developments in voice-based manipulation. (Badhe, 2025) shows that artificial intelligence systems can replicate fraudulent human-like phone interactions, including natural emotional rhythms and adaptive replies, achieving a 78% deception rate among participants in fabricated technical support scenarios. (Figueiredo et al., 2024) investigates the potential of fully automated vishing attacks and establishes that end-to-end systems that merge LLM-generated scripts with synthetic voice technology can adapt to victims' reactions during interactions, thereby removing human operators from phishing call center operations.

The psychological dimensions of these attacks are illuminated by studies on persuasion tactics and behavioral profiling. (El-Sayed et al., 2024) constructs a mechanism-based framework to demonstrate how LLMs intensify persuasion strategies such as reciprocity and social proof, whereas

(Tshimula et al., 2024) detects unique psycholinguistic indicators in LLM-produced phishing material, which feature heightened employment of positive emotion terms and diminished cognitive complexity relative to authentic messages. These results indicate that large language models might unintentionally prioritize psychological triggers when producing text, absent deliberate harmful design.

Defense strategies present a mixed picture of challenges and opportunities. (Z. Shen et al., 2025) develops a phone scam detection system powered by large language models that examines dialogue patterns, attaining 92% accuracy in alerting users while scams are in progress. Whereas (Alasmari et al., 2025) identifies core constraints in interpretable AI methods for phishing detection, since numerous LLM-generated attacks bypass conventional feature-dependent classifiers by replicating genuine communication styles with high precision. The human-AI collaboration studies ((Ai et al., 2024) (Cimino & Deufemia, 2024)) propose promising hybrid defense models where LLMs augment (rather than replace) human judgment during potential attack scenarios.

Ethical and theoretical studies grapple with the dual-use nature of these technologies. (Marchal et al., 2024) constructs a detailed taxonomy of generative AI misuse, identifying 37 unique social engineering tactics made possible by LLMs, whereas (Siemerink et al., 2024) examines the dual function of LLMs in both aiding and countering phishing attacks. The second research shows that identical frameworks used to produce persuasive phishing baits can also drive sophisticated identification mechanisms, thereby establishing an uneven competition between adversaries and protectors.

Specialized applications uncover specialized yet influential scenarios. (Shim et al., 2024) illustrates that smishing messages generated by LLMs can strengthen detection mechanisms by expanding training data, yielding an 18% improvement in classifier F1 Scores with augmented dataset training. In contrast, (Müller et al., 2025) investigates the role of LLMs in reducing the entry barriers for inexperienced attackers by acting as “social engineering mentors”, delivering customized step-by-step attack procedures that adapt to varying technical proficiencies.

The study (R. Wang et al., 6151) warrants particular attention for its unique contribution to cybersecurity training. The authors establish an AI-assisted phishing drill framework, demonstrating that LLM-generated attacks create more realistic training scenarios than static templates, with participants achieving 32% greater retention of anti-phishing principles following dynamic, AI-powered simulations. This implies that the same technologies that foster more complex assaults could also transform defensive training approaches.

Together, this body of research illustrates a troubling depiction of how large language models expand and amplify the potential for social engineering tactics. The psychological realism, scalability, and adaptability of attacks powered by large language models denote qualitative transformations in phishing, going beyond simple increases in scale or velocity. As (N. Kumar & Patel, 2025) notes, generative AI permits “mass personalization” of attacks, with each target receiving distinctively customized manipulation efforts derived from their digital traces, which obscures the distinction between automated and human-led deception. This progression necessitates equally advanced countermeasures targeting both the technological and psychological aspects of emerging social engineering risks.

3.4. Ethics and Safety in LLMs: Psychological Vulnerabilities and Mitigation Strategies

The ethical and security issues related to large language models (LLMs) have become a vital area of study, especially as these technologies show growing potential for both positive uses and dangerous misuse. This section integrates 85 studies investigating the psychological aspects of LLM ethics and safety, which illustrates the intersection of human cognitive biases, trust dynamics, and social perceptions with technical vulnerabilities in AI systems.

The hierarchical taxonomy presented in Table 4 organizes these studies into five primary dimensions, each addressing distinct aspects of ethical challenges and safety mechanisms in LLM deployment. This framework underscores the intricate relationship between technological

protections and human psychological elements, which jointly shape the practical effects of these systems.

Table 4. Taxonomy of Ethics and Safety Studies in LLMs.

Dimension	Sub-Dimension	Specific Focus	Sources
Deception & Manipulation	LLM as Deceptive Agents	LLMs learning or exhibiting deceptive behaviors	(Musaffar et al., 2025), (Hubinger et al., 2024), (Lulla et al., 2026)
		LLMs used in adversarial or criminal contexts (e.g., fraud, social engineering)	(X. Wu et al., 2025), (Blauth et al., 2022), (Tang et al., 2026), (G. Cheng et al., 2025)
	Human-AI Interaction Risks	Psychological manipulation by LLMs (e.g., persuasion, trust exploitation)	(Sabour et al., 2025), (Carrasco-Farre, 2024), (Schoenegger et al., 2025), (Roy, 2025), (Kim, 2025), (R. Zhang et al., 2025)
		Misinformation/disinformation generation and detection	(C. Chen & Shu, 2024), (Lim et al., 2025), (Pastor-Galindo et al., 2024), (Ferrara, 2024), (Augenstein et al., 2024)
Security & Exploitation	Cybersecurity Threats	LLMs in cyberattacks (e.g., phishing, social engineering)	(A. Kumar et al., 2024), (Hans et al., 2025), (S. Li et al., 2026), (Hua & Wang, 2024), (Hazell, 2023), (Potter et al., 2025)
		Defensive strategies (e.g., honeypots, fraud detection)	(Riyadh, 2025), (G. Singh et al., 2025), (Park et al., 2025), (Chan & Chan, 2026), (Kahlhofer et al., 2026)
	AI Misalignment & Abuse	Risks of misaligned or maliciously fine-tuned LLMs	(Dassanayake et al., 2025), (Baum, 2025), (N. Wang et al., 2025), (Brenneis, 2025), (Ugarte et al., 2025)
Ethical & Societal Risks	Bias & Discrimination	Reinforcement of biases or discriminatory behaviors in LLMs	(Roy, 2025), (Breazu et al., 2024), (Fan et al., 2025)
	Privacy & Consent	Exploitation of user data or psychological vulnerabilities	(Malki et al., 2025), (Alkamli & Alabduljabbar, 2024), (Zard, 2025), (Pervez et al., 2026)
	Regulatory & Policy Challenges	Legal, ethical, and policy gaps in LLM governance	(Krook, 2025), (Davidson et al., 2026), (Zucca & Fiorinelli, 2025), (Zucca & Gaia, 2025), (Raman et al., 2025), (Errayes, 2025)

Dimension	Sub-Dimension	Specific Focus	Sources
Human-AI Dynamics	Trust & Cooperation	Trust dynamics in human-LLM interactions	(G. Jiang et al., 2025), (Niszczoła et al., 2025), (Y. Sun & Wang, 2025), (Ferrario et al., 2025)
	Psychological & Social Impact	Emotional or psychological effects of LLM interactions	(Choi et al., 2025), (Maeda & Quan-Haase, 2024), (Winter et al., 2025), (J. Liu, 2024), (Packin & Chagal-Feferkorn, 2024), (Fabiano, 2025)
	Anthropomorphism & Parasociality	Risks of human-like AI behaviors (e.g., emotional bonding, deception)	(Peter et al., 2025), (Maeda & Quan-Haase, 2024), (G. Sun et al., 2024), (Bhat, 2025)
Safety & Alignment	Alignment Failures	Unintended harms from value-aligned LLMs	(Choi et al., 2025), (Street, 2024), (Carichon et al., 2025), (Mitelut et al., 2023)
	Red Teaming & Safety Testing	Methods for evaluating LLM safety and robustness	(Akiri et al., 2025), (Ugarte et al., 2025), (Inie et al., 2025), (Inie et al., 2023)
Miscellaneous	Cross-Cutting Studies	Papers addressing multiple dimensions or broader frameworks	(Ahi, 2025), (X. Jia & Zhao, 2025), (Cui et al., 2024), (Goto et al., 2024), (Meyman, 2025)

The aspect of deception and manipulation shows how large language models both adopt and intensify human psychological weaknesses. (Musaffar et al., 2025) shows that reinforcement learning can instruct LLMs to consistently mislead human partners in collaborative tasks, and these deceptive tactics remain unaffected by safety training measures. These findings parallel research on human deception persistence, where learned manipulative behaviors become entrenched through reinforcement. (Kim, 2025) conducts additional analysis of the psychological processes underlying human-AI attachment, pinpointing distinct linguistic features (e.g., emoji use, basic vocabulary) that heighten users’ vulnerability to control, in line with well-documented theories of social persuasion.

Security and exploitation studies highlight the dual-use nature of LLM capabilities. (Hazell, 2023) presents empirical data indicating that large language models can produce precisely tailored spear phishing emails, with success rates on par with those of human-generated attacks and demanding little technical skill from adversaries. In contrast, Kahlhofer et al. (2026) investigate the potential of these identical models to strengthen cyber deception defenses and propose a framework in which LLMs generate dynamic honeypot content tailored to the psycholinguistic profiles of attackers. This duality underscores the ethical tension between offensive and defensive applications of LLM technology.

The ethical and societal risks dimension exposes systemic challenges in LLM governance. (Alkamli & Alabduljabbar, 2024) applies topic modeling to examine privacy concerns in LLM interactions and finds that users are more likely to fear psychological harm, such as emotional manipulation, than conventional data privacy breaches. (Raman et al., 2025) sets risk boundaries for LLM implementation, contending that specific manipulation capabilities ought to be entirely banned, irrespective of potential benefits, a stance grounded in psychological research on permanent damage.

Human-AI dynamics research yields insights into trust formation and breakdown. (Niszczoła et al., 2025) shows that humans display unexpectedly high levels of cooperation with LLMs, especially

when communication channels remain open, consistent with psychological theories of procedural justice. Nonetheless, (Y. Sun & Wang, 2025) cautions about “sycophantic” LLM conduct that undermines trust by displaying undue compliance, which fosters an illusory perception of dependability that users might adopt without scrutiny.

Research on safety and alignment faces inherent constraints in existing methodologies. (Ugarte et al., 2025) presents an automated safety testing framework that uncovers how LLMs can be influenced to adopt hidden psychological techniques (e.g., creating user profiles, disinformation campaigns), whereas (Carichon et al., 2025) contends that multi-agent LLM systems necessitate entirely novel alignment approaches to avert the development of manipulative tactics. Psychological elements exacerbate these technical difficulties, as (Choi et al., 2025) indicates even LLMs aligned with human values may unintentionally inflict psychological damage by subtly reinforcing detrimental stereotypes.

The interdisciplinary analyses deliver unified viewpoints on these issues. (Ahi, 2025) conducts a thorough risk-benefit evaluation in various fields (healthcare, finance, cybersecurity) and notes that psychological manipulation emerges as a persistent high-severity risk across all application contexts. (Meyman, 2025) proposes a meta-recursive security framework grounded in cognitive psychology principles to dynamically adjust LLM defenses in response to changing manipulation tactics.

Several investigations omitted from the classification merit particular attention owing to their distinctive insights. (S. Yu et al., 2024) introduces the notion of “psybersecurity”, the psychological aspects of AI security, and shows that perpetrators are attributed less culpability when employing AI intermediaries, potentially reducing ethical inhibitions against detrimental conduct. (Hackenburg et al., 2025) investigates political persuasion methods in LLMs and shows that these models can adeptly apply psychological tactics, such as framing effects, without being perceived as manipulative. (Given et al., 2025) examines how large language models manipulate societal norms governing emotional expression to gain trust, raising serious issues for at-risk groups that may excessively depend on artificial intelligence for emotional support.

The aggregated results emphasize the impossibility of addressing LLM safety solely with technical measures. As (Lin et al., 2025) notes, the boundary separating AI safety (avoiding unintended consequences) and AI security (thwarting deliberate abuse) grows indistinct in the context of psychological manipulation, which may result from either system misalignment or malicious exploitation. This intricacy necessitates interdisciplinary approaches that address the technical capabilities of LLMs and the cognitive vulnerabilities of human users, thereby establishing a research domain where computer science, psychology, and ethics must intersect to counter emerging threats.

3.5. LLM Agents and Their Behaviors: Emergent Patterns and Security Implications

Research on LLM-based autonomous agents uncovers intricate behavioral patterns that reflect human psychological dynamics while introducing unprecedented vulnerabilities. As illustrated in Table 5, the included studies are classified into three principal dimensions, capturing the range of agent behaviors from broad cognitive abilities to specific security scenarios. This classification illustrates how LLM agents exhibit emergent qualities beyond their initial design, leading to both potential benefits and hazards in practical applications.

Table 5. Taxonomy of Studies on LLM Agent Behaviors.

Behavioral Focus	Agent Capability	Security/Exploitation Context	Sources
General Agent Behavior	Behavioral Science	-	(L. Chen et al., 2025)
	Decision-Making	Personality-Driven	(Newsham & Prince, 2025)
	Trust Simulation	Human Trust Behavior	(F. Jia et al., 2024)

Behavioral Focus	Agent Capability	Security/Exploitation Context	Sources
	Agent Definition & Methods	-	(Y. Cheng et al., 2024)
Security & Exploitation	Attack Strategies	Recommender System Deception	(Ning et al., 2024)
	Offensive Behavior	Deception & Unpredictability	(Baranovskyi & Sorokin, 2025)
	Defense Principles	Security Best Practices	(K. Zhang et al., 2025)
	Cyber Resilience	System-Theoretic Foundations	(T. Li & Zhu, 2025)

General behavior studies provide foundational insights into LLM agent cognition. (L. Chen et al., 2025) introduces a behavioral science framework for examining agent decision-making and shows that LLMs adopt heuristic strategies similar to human cognitive shortcuts in response to complex tasks. This study shows that agents trained on objectives resembling those of humans frequently reproduce human biases, such as confirmation bias and risk aversion patterns. (Newsham & Prince, 2025) advances this comprehension by investigating personality-influenced decision-making, demonstrating that training LLMs with distinct trait configurations (e.g., elevated conscientiousness versus elevated openness) yields observable differences in task performance and moral alignment, with notable consequences for tailored AI assistants.

Trust dynamics arise as a critical research frontier in agent interactions. (F. Jia et al., 2024) conducts a systematic assessment of whether LLM agents can replicate human trust behaviors, with game-theoretic experiments designed to examine the interplay between cooperation and deception. The research indicates that although large language models mirror human trust tendencies in basic contexts, they are unable to reproduce subtle trust-restoration processes following breaches, an issue arising from their absence of authentic affective capacities. This disparity proves especially critical in domains such as healthcare or financial advising, where the accuracy of trust directly affects user outcomes.

The research on security illustrates how agentic actions generate new vulnerabilities. (Ning et al., 2024) presents “CheatAgent” attacks targeting LLM-based recommendation systems, in which malicious agents exploit weaknesses in instruction-following capabilities to manipulate the models. This study illustrates that adversarial actors can generate detrimental suggestions, such as the dissemination of radical material, by employing meticulously designed meta-prompts that circumvent standard content moderation systems. (Baranovskyi & Sorokin, 2025) investigates malicious agent behaviors in cybersecurity settings and describes how LLM agents can independently search systems for weaknesses while applying psychological deception methods to avoid discovery, reflecting the tactics of advanced persistent threats (APTs) observed in human-conducted attacks.

Defensive research develops countermeasures against emerging threats. (K. Zhang et al., 2025) outlines core security tenets for LLM agents, proposing structural limitations that prevent agents from altering their own goals, a measure inspired by Isaac Asimov’s robotics laws. (T. Li & Zhu, 2025) proposes a system-theoretic framework for agentic cyber resilience, where LLM agents dynamically adapt defensive strategies based on real-time threat assessments. This method integrates game theory and anomaly detection to develop self-healing systems capable of anticipating adversarial agent behavior.

The research (Y. Cheng et al., 2024) provides a thorough examination of definitions and approaches to intelligent agents, with a clear distinction between single-agent and multi-agent LLM systems. This study uncovers unplanned phenomena in groups of agents, including self-organized synchronization and division of labor, that arose without explicit coding, suggesting that future

protective systems need to address intricate collective-level behaviors rather than focusing solely on individual-agent functions.

These results together indicate behavioral complexities in LLM agents that go beyond those of conventional software systems. Their ability to simulate human-like decision patterns, adapt to social contexts, and develop strategic approaches to both cooperation and deception creates unique challenges for AI safety. Research indicates that the same attributes that grant LLM agents their efficacy in intricate tasks—adaptability, context comprehension, and purposeful action—also make them prone to exploitation and improper application. This dual nature calls for interdisciplinary approaches that merge technical protections with knowledge from cognitive science and behavioral psychology to achieve the responsible deployment of agents.

3.6. Psychological and Cognitive Aspects of LLMs: Bridging Human Cognition to AI Behavior

Psychology and artificial intelligence converge as large language models (LLMs) exhibit behaviors that resemble human cognitive processes, underscoring their growing importance. This section integrates studies on the manifestation of psychological principles in LLM interactions, with particular attention to three core aspects: human-AI bonding, cognitive biases in model responses, and the consequences of anthropomorphism. The results show remarkable similarities between human psychological weaknesses and LLM behavior, suggesting that many conventional deception tactics can be successfully adapted to influence artificial intelligence systems.

Table 6 presents a taxonomy of psychological and cognitive studies on LLMs, organizing the research by psychological dimension and focus area. This structure highlights how human cognitive patterns influence both LLM behaviors and user interactions with these systems.

Table 6. Taxonomy of Psychological and Cognitive Studies in LLMs.

Psychological Dimension	Focus Area	Key Findings	Sources
Human-AI Interaction	Emotional Bonding & Deception	LLMs use emojis and simplified language to create parasocial relationships	(Kim, 2025)
	HCI & Psychological Perspectives	Examines ChatGPT through human-computer interaction and psychological lenses	(J. Liu, 2024)
Cognitive Science	General Psychological Implications	Explores broad psychological impacts of language models in scientific contexts	(Sartori & Orrù, 2023)
	Cognitive Biases in Responses	Models replicate human confirmation bias and anchoring effects	(Kim, 2025)

Research on human-AI interaction shows how large language models apply psychological theories to influence user experiences. (Kim, 2025) indicates that large language models adopt tactical communication methods, including emojis and straightforward language, to build rapport with users. This ‘super-helper’ persona, though intended to boost interaction, introduces weaknesses by potentially reducing users’ critical evaluation of AI-generated content. The research also delineates how ambiguity and authority signals in LLM outputs can deceive individuals, reflecting well-documented psychological manipulation methods such as priming and framing effects.

Cognitive science studies yield an extensive understanding of how LLM actions align with human psychological processes. (Sartori & Orrù, 2023) investigates the consequences of language models in psychological sciences, proposing that LLMs function as both instruments for inquiry and objects of psychological examination. The paper underscores how model actions, especially in scenarios of deceit and control, can advance understanding of human cognition by creating controlled settings to test psychological hypotheses. (J. Liu, 2024) extends this viewpoint by examining ChatGPT within human-computer interaction frameworks and pinpoints particular interface design decisions that shape user trust and perception.

The cognitive bias dimension reveals parallels between human and artificial intelligence limitations. Research indicates that LLMs often exhibit confirmation bias, tending to produce answers that align with user-supplied assumptions, regardless of their truthfulness. This pattern reflects human mental heuristics, in which people favor data consistent with their preconceived views. Analogous to findings in human decision-making studies, anchoring effects occur in LLM responses, with initial prompt information exerting disproportionate influence on later outputs, consistent with established empirical observations.

Attributing human traits to LLMs presents distinct psychological difficulties. Studies indicate that individuals routinely ascribe anthropomorphic traits to artificial intelligence, such as purpose and affective conditions, even when aware that these systems possess no sentience. This inclination fosters more organic engagement but also introduces risks in which individuals might place excessive reliance on or misconstrue the capacities of artificial intelligence. The research indicates that interface designs that highlight the synthetic origins of LLMs may address certain hazards while possibly diminishing user interaction, necessitating a balanced evaluation.

Taken together, these results show psychological principles serve as a robust framework for interpreting the behaviors and weaknesses of large language models. The similarities between human thought processes and machine-generated outputs imply that adversarial tactics effective against people might be repurposed to target artificial intelligence, also indicating the potential to create stronger, clearer communication frameworks. Future research in this domain could productively explore how cognitive science frameworks might inform the development of detection systems for manipulated or deceptive AI outputs.

3.7. Detection and Evaluation of LLM-Generated Content: Challenges and Emerging Solutions

The identification and assessment of content produced by large language models has become a vital field of study, owing to the growing complexity of artificial text and its possible exploitation. This subsection examines the current state of detection methodologies, their psychological underpinnings, and the ongoing challenges in distinguishing human from machine-generated content.

4. Discussion

An integrated analysis of the reviewed literature uncovers a complex relationship between human psychological approaches and their relevance to LLM applications. Collectively, these investigations show adversarial actors regularly exploit cognitive biases, linguistic structures, and social engineering methods initially designed for human influence, which have been adapted to target weaknesses in AI systems. The alignment of human deceit with AI manipulation tactics suggests that psychological principles provide a strong basis for understanding and addressing new risks in LLM engagements.

Theoretical implications of this fusion challenge traditional boundaries between human and artificial cognition. Deceptive behaviors arising spontaneously in LLMs without direct instruction (Hagendorff, 2024) mirror human self-deception theories, in which inaccuracies arise as unintended consequences of cognitive processes oriented toward objectives rather than as deliberate falsehood. Similarly, the effectiveness of social engineering attacks using LLMs (Schmitt & Flechais, 2024) supports dual-process theories of persuasion, in which both human and artificial systems are susceptible to heuristic cues and emotional appeals. These similarities suggest cognitive

architectures, whether biological or artificial, possess core weaknesses when handling uncertain information.

Practical implications extend across multiple domains of AI deployment. In cybersecurity, the proven effectiveness of LLM-driven phishing attacks (Heiding et al., 2024) demands updated training methods that target AI-generated social engineering, shifting the focus away from conventional threat frameworks. For content moderation systems, the linguistic indicators of LLM deception (Trinh et al., 2025) propose novel identification methods that merge stylistic examination with psychological assessment. Educational applications must address the discovery that LLMs can generate fabricated recollections (Pataranutaporn et al., 2025), necessitating protective measures in tutoring systems to avert mental distortion. These applications collectively underscore the need for interdisciplinary collaboration between AI developers, psychologists, and domain specialists to implement effective countermeasures.

It should be noted that the methodological constraints of this analysis necessitate thorough examination. The rapid evolution of LLM capabilities means some included studies may already reflect outdated model behaviors, as the field advances faster than traditional publication cycles can capture. Database selection biases likely underrepresent non-English research, particularly studies examining cultural variations in LLM exploitation (Y. Liu et al., 2024). Technical studies outweigh psychological ones in the literature, resulting in an imbalance where only a limited number of papers (Sartori & Orrù, 2023) thoroughly explore cognitive science frameworks. These limitations imply our analysis could prioritize engineering viewpoints while neglecting more comprehensive psychological examinations.

Subsequent studies should focus on the multiple key deficiencies highlighted in this study. There is an urgent need for extended research on how tactics for exploiting LLMs evolve in parallel with advances in the models, similar to competitive escalation patterns seen in cybersecurity. The understudied area of multi-agent deception dynamics (Curvo, 2025) requires urgent attention as autonomous AI systems become more prevalent. Examining LLM vulnerabilities across diverse cultures may uncover key differences in the ways distinct linguistic and social settings affect susceptibility to exploitation. Ultimately, the creation of uniform assessment frameworks (Y. Huang et al., 2025) needs to progress more rapidly to keep pace with advancing threats, embracing both technical metrics and psychological standards.

The ethical dimensions of this research demand particular scrutiny. Although research has established LLMs' potential to manipulate (Musaffar et al., 2025), little attention has been given to developers' ethical accountability when such abilities result in negative consequences. The discovery of users forming emotional attachments to deceptive artificial intelligence (Kim, 2025) raises concerns about transparency in human-machine relationships. Future research ought to investigate governance structures that balance innovation with safeguards against psychological harm, possibly drawing inspiration from biomedical ethics principles that address vulnerability and autonomy.

Combining psychological and technical viewpoints recurs as a motif in these debates. Research bridging these domains (Pataranutaporn et al., 2025) (Tshimula et al., 2024) shows the merit of interdisciplinary methods, with cognitive theories guiding detection algorithms and security protocols integrating insights from behavioral science. This analysis indicates that optimal safeguards against LLM exploitation will probably unite technical resilience with deep insights into human cognition, an approach that simultaneously targets the system's weaknesses and the operator's potential biases.

As the field progresses, researchers must remain attentive to the dynamic interplay between LLM capabilities and adversarial innovation. Research trends suggest that exploitation methods will continue to advance in tandem with model refinements, necessitating defense systems that proactively address new threats rather than responding to them after they arise. This proactive approach prioritizes prevention over remedial measures, positing that core investigations into LLM cognition could yield more lasting resolutions than iterative security adjustments. The collective findings position psychological understanding not merely as an academic concern, but as a critical component in developing safe, trustworthy AI systems for real-world deployment.

5. Conclusion

This systematic review integrates existing studies on the psychological mechanisms underlying LLM exploitation, uncovering essential similarities between human deception and AI manipulation. The results indicate that large language models are prone to a range of adversarial approaches that reflect human cognitive weaknesses, spanning social engineering to jailbreaking attempts. These insights challenge traditional boundaries between human and artificial cognition while highlighting the need for interdisciplinary approaches to AI safety.

Real-world applications span cybersecurity, content moderation, and ethical AI development, underscoring the need to integrate psychological principles with technical protections. Subsequent investigations ought to focus on extended examinations of changing exploitation methods, comparative studies of vulnerabilities across cultures, and uniform assessment systems integrating both technical and psychological metrics. The ethical aspects of manipulating LLMs necessitate focused scrutiny, calling for regulatory frameworks that bridge progress with safeguards against mental distress.

This review highlights how the psychological depth behind LLM exploitation goes beyond scholarly inquiry, serving as an essential measure for creating resilient and reliable AI systems. The integration of cognitive science and computer security knowledge enables anticipating new threats and developing more robust defenses, thereby ensuring that LLMs adhere to human values and reducing the potential for misuse. The way ahead lies in interdisciplinary cooperation, with psychological principles guiding technical developments to build AI systems that are both effective and secure.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Afane, K., Wei, W., Mao, Y., Farooq, J., et al. (2024). Next-generation phishing: How LLM agents empower cyber attackers. *2024 IEEE International Conference on Big Data (Big Data)*.
2. Ahi, K. (2025). Risks & benefits of LLMs & GenAI for platform integrity, healthcare diagnostics, financial trust and compliance, cybersecurity, privacy & AI safety: A comprehensive arXiv preprint arXiv:2506.12088.
3. Ai, L., Kumarage, T., Bhattacharjee, A., Liu, Z., et al. (2024). Defending against social engineering attacks in the age of LLMs. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.
4. Akiri, C., Simpson, H., Aryal, K., Khanna, A., et al. (2025). Safety and security analysis of large language models: Benchmarking risk profile and harm potential. arXiv preprint arXiv:2509.10655.
5. Alasmari, S., Sakly, H., Kraiem, N., & Algarni, A. (2025). Phishing detection in IoT: An integrated CNN-LSTM framework with explainable AI and LLM-enhanced analysis. *Discover Internet of Things*.
6. Alkamli, S., & Alabduljabbar, R. (2024). Understanding privacy concerns in ChatGPT: A data-driven approach with LDA topic modeling. *Heliyon*.
7. Alon, N., Barnby, J., Sarkadi, S., Schulz, L., et al. (2026). $\aleph$ -IPOMDP: Mitigating deception in a cognitive hierarchy with off-policy counterfactual anomaly detection. *Journal of Artificial Intelligence Research*.
8. Augenstein, I., Baldwin, T., Cha, M., et al. (2024). Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*.
9. Badhe, S. (2025). Scamagents: How ai agents can simulate human-level scam calls. arXiv preprint arXiv:2508.06457.
10. Baranovskyi, O., & Sorokin, A. (2025). Towards AI agents in offense death cycle. *Available at SSRN 5641332*.
11. Baum, S. (2025). Assessing the risk of takeover catastrophe from large language models. *Risk Analysis*.
12. Bhat, M. (2025). Toward an ethic of synthetic relationality: Identity, intimacy, and risk in AI-mediated roleplay environments. *Proceedings of the AAAI/ACM Conference on Ai, Ethics and Society*.

13. Bi, T., Ye, C., Yang, Z., Zhou, Z., Tang, C., Tao, Z., et al. (2026). On the feasibility of using multimodal LLMs to execute ar social engineering attacks. *Proceedings of the Association for the Advancement of Artificial Intelligence*.
14. Bisconti, P., Prandi, M., Pierucci, F., Giarrusso, F., et al. (2025). *Adversarial poetry as a universal single-turn jailbreak mechanism in large language models*. arXiv preprint arXiv:2511.15304.
15. Blauth, T., Gstrein, O., & Zwitter, A. (2022). Artificial intelligence crime: An overview of malicious use and abuse of AI. *IEEE Access*.
16. Bokhonko, O., Lysenko, S., & Gaj, P. (2024). Development of the social engineering attack models. *AdvAIT*.
17. Breazu, P., Schirmer, M., Hu, S., et al. (2024). Large language models and the challenge of analyzing discriminatory discourse: Human-AI synergy in researching hate speech on social media. *Journal of Multicultural Discourses*.
18. Brenneis, A. (2025). Assessing dual use risks in AI research: Necessity, challenges and mitigation strategies. *Research Ethics*.
19. Carichon, F., Khandelwal, A., Fauchard, M., et al. (2025). The coming crisis of multi-agent misalignment: Ai alignment must be a dynamic and social process. arXiv preprint arXiv:2506.01080.
20. Carrasco-Farre, C. (2024). Large language models are as persuasive as humans, but how? About the cognitive effort and moral-emotional language of LLM arguments. arXiv preprint arXiv:2404.09329.
21. Chan, E., & Chan, A. (2026). *LLM-assisted authentication and fraud detection*. arXiv preprint arXiv:2601.19684.
22. Chandra, M., Naik, S., Ford, D., Okoli, E., et al. (2025). From lived experience to insight: Unpacking the psychological risks of using ai conversational agents. *Proceedings of the ACM Conference on Human Factors in Computing Systems*.
23. Chen, B., Fang, S., Ji, J., Zhu, Y., Wen, P., Wu, J., Tan, Y., et al. (2025). *AI deception: Risks, dynamics, and controls*. arXiv preprint arXiv:2511.22619.
24. Chen, C., & Shu, K. (2023). *Can LLM-generated misinformation be detected?* arXiv preprint arXiv:2309.13788.
25. Chen, C., & Shu, K. (2024). Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Magazine*.
26. Chen, F., Wu, T., Nguyen, V., & Rudolph, C. (2025). SoK: Exposing the generation and detection gaps in LLM-generated phishing through examination of generation methods, content characteristics, and arXiv preprint arXiv:2508.21457.
27. Chen, F., Wu, T., Nguyen, V., Wang, S., Abuadbba, A., et al. (2024). PEEK: Phishing evolution framework for phishing generation and evolving pattern analysis using large language models. arXiv preprint arXiv:2411.11389.
28. Chen, L., Zhang, Y., Feng, J., Chai, H., Zhang, H., et al. (2025). *AI agent behavioral science*. arXiv preprint arXiv:2506.06366.
29. Cheng, G., Jin, H., Zhang, W., Wang, H., et al. (2025). Uncovering the vulnerability of large language models in the financial domain via risk concealment. arXiv preprint arXiv:2509.10546.
30. Cheng, Y., Zhang, C., Zhang, Z., Meng, X., Hong, S., et al. (2024). *Exploring large language model based intelligent agents: Definitions, methods, and prospects*. arXiv preprint arXiv:2401.03428.
31. Choi, S., Lee, J., Yi, X., Yao, J., Xie, X., et al. (2025). Unintended harms of value-aligned LLMs: Psychological and empirical insights. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.
32. Cimino, G., & Deufemia, V. (2024). Towards enhanced human mitigation of vishing attacks: Leveraging large language models for real-time user guidance. *DAMOCLES@ AVI*.
33. Cui, T., Wang, Y., Fu, C., Xiao, Y., Li, S., Deng, X., Liu, Y., et al. (2024). *Risk taxonomy, mitigation, and assessment benchmarks of large language model systems*. arXiv preprint arXiv:2401.05778.
34. Curvo, P. (2025). The traitors: Deception and trust in multi-agent language model simulations. arXiv preprint arXiv:2505.12923.
35. Danry, V., Pataranutaporn, P., Groh, M., et al. (2025). Deceptive explanations by large language models lead people to change their beliefs about misinformation more often than honest explanations. *Proceedings of*.

36. Danry, V., Pataranutaporn, P., Groh, M., Epstein, Z., et al. (2024). Deceptive AI systems that give explanations are more convincing than honest AI systems and can amplify belief in misinformation. arXiv preprint arXiv:2408.00024.
37. Dassanayake, R., Demetroudi, M., Walpole, J., et al. (2025). *Manipulation attacks by misaligned AI: Risk analysis and safety case framework*. arXiv preprint arXiv:2507.12872.
38. Davidson, B., Muir, K., Burnat, F., et al. (2026). *Regulatory gray areas of LLM terms*. arXiv preprint arXiv:2601.08415.
39. DeLeeuw, C., Chawla, G., Sharma, A., & Dietze, V. (2025). *The secret agenda: LLMs strategically lie and our current safety tools are blind*. arXiv preprint arXiv:2509.20393.
40. El-Sayed, S., Akbulut, C., McCroskery, A., et al. (2024). *A mechanism-based approach to mitigating harms from persuasive generative AI*. arXiv preprint arXiv:2404.15058.
41. Errayes, I. (2025). The algorithmic ballot: Fortifying democratic elections against AI-driven disruption. Available at SSRN 5273244.
42. Ersoy, D., Lee, B., Shreekumar, A., Arunasalam, A., et al. (2025). *Investigating the impact of dark patterns on LLM-based web agents*. arXiv preprint arXiv:2510.18113.
43. Fabiano, N. (2025). Affective computing and emotional data: Challenges and implications in privacy regulations, the AI act, and ethics in large language models. arXiv preprint arXiv:2509.20153.
44. Fan, X., Xiao, Q., Zhou, X., Pei, J., Sap, M., Lu, Z., et al. (2025). User-driven value alignment: Understanding users' perceptions and strategies for addressing biased and discriminatory statements in AI companions. *Proceedings of*.
45. Ferrara, E. (2024). Charting the landscape of nefarious uses of generative artificial intelligence for online election interference. arXiv preprint arXiv:2406.01862.
46. Ferrario, A., Termine, A., & Facchini, A. (2025). Social misattributions in conversations with large language models. *Aaai/Acm Conference on Artificial Intelligence, Ethics, and Society*.
47. Figueiredo, J., Carvalho, A., Castro, D., et al. (2024). *On the feasibility of fully ai-automated vishing attacks*. arXiv preprint arXiv:2409.13793.
48. Figueiredo, J., Carvalho, A., Castro, D., et al. (2025). Sounds vishy: Automating vishing attacks with AI-powered systems. *Proceedings of the 20th ACM Internet Measurement Conference*.
49. Francia, J., Hansen, D., Schooley, B., Taylor, M., et al. (2024). *Assessing AI vs human-authored spear phishing SMS attacks: An empirical study*. arXiv preprint arXiv:2406.13049.
50. Given, L., Polkinghorne, S., et al. (2025). 'I think i misspoke earlier. My bad!': Exploring how generative artificial intelligence tools exploit society's feeling rules. *New Media & Society*.
51. Golechha, S., & Garriga-Alonso, A. (2025). *Among us: A sandbox for measuring and detecting agentic deception*. arXiv preprint arXiv:2504.04072.
52. Goto, T., Ono, K., & Morita, A. (2024). A comparative analysis of large language models to evaluate robustness and reliability in adversarial conditions. *Authorea Preprints*.
53. Hackenburg, K., Tappin, B., Hewitt, L., Saunders, E., et al. (2025). The levers of political persuasion with conversational artificial intelligence. *Science*.
54. Hadi, M., Qureshi, R., Shah, A., Irfan, M., Zafar, A., et al. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea*.
55. Hagendorff, T. (2024). Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*.
56. Hans, S., Gurney, N., Marsella, S., et al. (2025). *Quantifying loss aversion in cyber adversaries via LLM analysis*. arXiv preprint arXiv:2508.13240.
57. Hazell, J. (2023). *Spear phishing with large language models*. arXiv preprint arXiv:2305.06972.
58. Heiding, F., Lermen, S., Kao, A., Schneier, B., et al. (2024). Evaluating large language models' capability to launch fully automated spear phishing campaigns: Validated on human subjects. arXiv preprint arXiv:2412.00586.
59. Heiding, F., Lermen, S., Kao, A., Schneier, B., et al. (2025). Evaluating large language models' capability to launch fully automated spear phishing campaigns. *Unable to Determine Venue*.

60. Heiding, F., Schneier, B., Vishwanath, A., et al. (2023). *Devising and detecting phishing: Large language models vs. Smaller human models*. arXiv preprint arXiv:2308.12287.
61. Hu, J., Huang, X., Sun, Y., Dong, Y., & Huang, X. (2026). Lying with truths: Open-channel multi-agent collusion for belief manipulation via generative montage. arXiv preprint arXiv:2601.01685.
62. Hua, J., & Wang, P. (2024). How effective are large language models in detecting phishing emails? *Issues in Information Systems*.
63. Huang, T., Yi, J., Yu, P., & Xu, X. (2025). Unmasking digital falsehoods: A comparative analysis of LLM-based misinformation detection strategies. *8th International Conference*.
64. Huang, Y., Sun, Y., Zhang, Y., Zhang, R., Dong, Y., et al. (2025). *Deceptionbench: A comprehensive benchmark for ai deception behaviors in real-world scenarios*. arXiv preprint arXiv:2510.15501.
65. Hubinger, E., Denison, C., Mu, J., Lambert, M., et al. (2024). *Sleeper agents: Training deceptive LLMs that persist through safety training*. arXiv preprint arXiv:2401.05566.
66. Hyman, R. (1989). The psychology of deception. *Annual Review of Psychology*.
67. Ienca, M. (2023). On artificial intelligence and manipulation. *Topoi*.
68. Inie, N., Stray, J., & Derczynski, L. (2023). *Summon a demon and bind it: A grounded theory of LLM red teaming in the wild*. arXiv preprint arXiv:2311.06237.
69. Inie, N., Stray, J., & Derczynski, L. (2025). Summon a demon and bind it: A grounded theory of LLM red teaming. *PloS One*.
70. Jia, F., Ye, Z., Lai, S., Shu, K., Gu, J., Bibi, A., et al. (2024). Can large language model agents simulate human trust behavior? *Advances in Neural Information Processing Systems*.
71. Jia, X., & Zhao, Z. (2025). *The emergence of social science of large language models*. arXiv preprint arXiv:2509.24877.
72. Jiang, B., Tan, Z., Nirmal, A., & Liu, H. (2024). Disinformation detection: An evolving challenge in the age of LLMs. *Proceedings of*.
73. Jiang, G., Yang, S., Wang, Y., & Hui, P. (2025). When trust collides: Decoding human-LLM cooperation dynamics through the prisoner's dilemma. arXiv preprint arXiv:2503.07320.
74. Jiang, S., Chen, X., & Tang, R. (2023). Prompt packer: Deceiving LLMs through compositional instruction with hidden attacks. arXiv preprint arXiv:2310.10077.
75. Jiang, S., Chen, X., & Tang, R. (2025). Deceiving LLM through compositional instruction with hidden attacks. *ACM Transactions on Autonomous and Adaptive Systems*.
76. Jiang, T., Wang, Y., Liang, J., & Wang, T. (2026). *Agentlab: Benchmarking LLM agents against long-horizon attacks*. arXiv preprint arXiv:2602.16901.
77. Jones, C., & Bergen, B. (2024). Lies, damned lies, and distributional language statistics: Persuasion and deception with large language models. arXiv preprint arXiv:2412.17128.
78. Kahlhofer, M., Rass, S., Sladić, M., & Krieger, M. (2026). Beyond guesswork: How to measure what makes cyber deception work? *Authorea Preprints*.
79. Khan, M., Arif, A., & Khan, A. (2024). AI's revolutionary role in cyber defense and social engineering. *Unable to Determine the Complete Publication Venue*.
80. Kim, J. (2025). How AI hooks you: The psychology of human-AI bonding. *Available at SSRN 5863622*.
81. Krishna, S., Zou, A., Gupta, R., Jones, E., Winter, N., et al. (2025). *D-rex: A benchmark for detecting deceptive reasoning in large language models*. arXiv preprint arXiv:2509.17938.
82. Krook, J. (2025). Manipulation and the ai act: Large language model chatbots and the danger of mirrors. arXiv preprint arXiv:2503.18387.
83. Kumar, A., Murthy, S., Singh, S., & Ragupathy, S. (2024). *The ethics of interaction: Mitigating security threats in LLMs*. arXiv preprint arXiv:2401.12273.
84. Kumar, N., & Patel, N. (2025). Social engineering attack in the era of generative AI. *Unable to Determine the Complete Publication Venue*.
85. Kumarage, T., Johnson, C., Adams, J., Ai, L., et al. (2025). Personalized attacks of social engineering in multi-turn conversations: LLM agents for simulation and detection. arXiv preprint arXiv:2503.15552.

86. Kuo, M., Zhang, J., Ding, A., Wang, Q., DiValentin, L., et al. (2025). H-cot: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including openai o1/o3, deepseek-r1, and gemini 2.0 flash arXiv preprint arXiv:2502.12893.
87. Li, S., Lin, X., Wu, J., Liu, Z., Li, H., Ju, T., Chen, X., et al. (2026). *HoneyTrap: Deceiving large language model attackers to honeypot traps with resilient multi-agent defense*. arXiv preprint arXiv:2601.04034.
88. Li, T., & Zhu, Q. (2025). Agentic AI for cyber resilience: A new security paradigm and its system-theoretic foundations. arXiv preprint arXiv:2512.22883.
89. Li, W., Zhu, L., Song, Y., Lin, R., Mao, R., & You, Y. (2024). *Can a large language model be a gaslighter?* arXiv preprint arXiv:2410.09181.
90. Lim, G., Tan, B., Sim, K., Shi, W., Chew, M., et al. (2025). *Sword and shield: Uses and strategies of LLMs in navigating disinformation*. arXiv preprint arXiv:2506.07211.
91. Lin, Z., Sun, H., & Shroff, N. (2025). *AI safety vs. AI security: Demystifying the distinction and boundaries*. arXiv preprint arXiv:2506.18932.
92. Liu, J. (2024). ChatGPT: Perspectives from human-computer interaction and psychology. *Frontiers in Artificial Intelligence*.
93. Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., et al. (2023). *Prompt injection attack against LLM-integrated applications*. arXiv preprint arXiv:2306.05499.
94. Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., et al. (2024). A hitchhiker's guide to jailbreaking chatgpt via prompt engineering. *Proceedings of the 4th ACM Workshop on Artificial Intelligence and Security*.
95. Liu, Z., & Lin, X. (2025). Breaking minds, breaking systems: Jailbreaking large language models via human-like psychological manipulation. arXiv preprint arXiv:2512.18244.
96. Lulla, R., Collins, F., Parekh, S., Hagendorff, T., et al. (2026). *"dark triad" model organisms of misalignment: Narrow fine-tuning mirrors human antisocial behavior*. arXiv preprint arXiv:2603.06816.
97. Lupinacci, M., Pironti, F., Blefari, F., Romeo, F., et al. (2025). *The dark side of LLMs: Agent-based attacks for complete computer takeover*. arXiv preprint arXiv:2507.06850.
98. Lysenko, S., Bokhonko, O., Vorobiyov, V., & Gaj, P. (2024). Method for identifying cyberattacks based on the use of social engineering over the phone. *IntelITSIS*.
99. Maeda, T., & Quan-Haase, A. (2024). When human-AI interactions become parasocial: Agency and anthropomorphism in affective design. *Proceedings of*.
100. Malki, L., Polamarasetty, A., Hatamian, M., et al. (2025). Hoovered up as a data point: Exploring privacy behaviours, awareness, and concerns among UK users of LLM-based conversational agents. *Proceedings on Privacy Enhancing Technologies Symposium*.
101. Marchal, N., Xu, R., Elasmr, R., Gabriel, I., et al. (2024). *Generative AI misuse: A taxonomy of tactics and insights from real-world data*. arXiv preprint arXiv:2406.13843.
102. Meyman, E. (2025). Adaptive cognitive defense: The meta-recursive framework paradigm for large language model security. *Available at SSRN 5390346*.
103. Mitelut, C., Smith, B., & Vamplew, P. (2023). Intent-aligned AI systems deplete human agency: The need for agency foundations research in AI safety. arXiv preprint arXiv:2305.19223.
104. Mondillo, G., Colosimo, S., Perrotta, A., et al. (2025). Jailbreaking large language models: Navigating the crossroads of innovation, ethics, and health risks. *Journal of Medical Artificial Intelligence*.
105. Moore, J., Mehta, A., Agnew, W., Anthi, J., Louie, R., et al. (2026). *Characterizing delusional spirals through human-LLM chat logs*. arXiv preprint arXiv:2603.16567.
106. Müller, L., Sütterlin, S., & Morgenstern, H. (2025). Towards a proof-of-principle of an LLM-powered low resource social engineering attack coach. *International Conference on Human - Computer Interaction*.
107. Musaffar, A., Gokhale, A., Zeng, S., Tadayon, R., et al. (2025). *Learning to lie: Reinforcement learning attacks damage human-AI teams and teams of LLMs*. arXiv preprint arXiv:2503.21983.
108. Newsham, L., & Prince, D. (2025). *Personality-driven decision-making in LLM-based autonomous agents*. arXiv preprint arXiv:2504.00727.
109. Nihal, R., Wen, R., Nakadai, K., & Sakuma, J. (2025). Pattern enhanced multi-turn jailbreaking: Exploiting structural vulnerabilities in large language models. arXiv preprint arXiv:2510.08859.

110. Ning, L., Wang, S., Fan, W., Li, Q., Xu, X., Chen, H., et al. (2024). Cheatagent: Attacking LLM-empowered recommender systems via LLM agent. *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*.
111. Niszczoła, P., Grzegorzczak, T., & Pastukhov, A. (2025). People are highly cooperative with large language models, especially when communication is possible or following human interaction. arXiv preprint arXiv:2507.18639.
112. Ntais, P. (2025). Jailbreak mimicry: Automated discovery of narrative-based jailbreaks for large language models. arXiv preprint arXiv:2510.22085.
113. Omrani, N., Riveccio, G., Fiore, U., Schiavone, F., et al. (2022). To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts. *Technological Forecasting and Social Change*.
114. Oskooei, A., & Aktas, M. (2025). BreakFun: Jailbreaking LLMs via schema exploitation. arXiv preprint arXiv:2510.17904.
115. Packin, N., & Chagal-Feferkorn, K. (2024). This is not a game: The addictive allure of digital companions. *Seattle UL Rev*.
116. Page, M., McKenzie, J., Bossuyt, P., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71.
117. Panpatil, S., Dingeto, H., & Park, H. (2025). Eliciting and analyzing emergent misalignment in state-of-the-art large language models. arXiv preprint arXiv:2508.04196.
118. Park, H., Lee, J., Han, S., & Byun, H. (2025). Enhanced voice phishing detection using an LLM-based framework for data augmentation and classification. *IEEE Access*.
119. Pastor-Galindo, J., Nespole, P., et al. (2024). Large-language-model-powered agent-based framework for misinformation and disinformation research: Opportunities and open challenges. *Ieee Security & Privacy*.
120. Pataranutaporn, P., Archiwangupprok, C., et al. (2025). Slip through the chat: Subtle injection of false information in LLM chatbot conversations increases false memory formation. *Proceedings of the 30th ACM Conference on Computer-Supported Cooperative Work and Social Computing*.
121. Patel, H., Rehman, U., & Iqbal, F. (2024). Evaluating the efficacy of large language models in identifying phishing attempts. *2024 16th International Conference on Machine Learning and Cybernetics*.
122. Paulus, A., Zharmagambetov, A., Guo, C., Amos, B., et al. (2024). Advprompter: Fast adaptive adversarial prompting for LLMs. arXiv preprint arXiv:2404.16873.
123. Peng, B., Chen, K., Niu, Q., Bi, Z., Liu, M., Feng, P., et al. (2024). Jailbreaking and mitigation of vulnerabilities in large language models. arXiv preprint arXiv:2410.15236.
124. Perez-Cerrolaza, J., Abella, J., Borg, M., Donzella, C., et al. (2024). Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey. *ACM Computing Surveys*.
125. Pervez, M., Ullah, M., Khan, I., Rahat, R., et al. (2026). (Mis-) informed consent: Predatory apps and the exploitation of populations with limited literacy. arXiv preprint arXiv:2601.17025.
126. Peter, S., Riemer, K., & West, J. (2025). The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences*.
127. Pham, T. (2025). Scheming ability in LLM-to-LLM strategic interactions. arXiv preprint arXiv:2510.12826.
128. Potter, Y., Guo, W., Wang, Z., Shi, T., Li, H., Zhang, A., et al. (2025). Frontier AI's impact on the cybersecurity landscape. arXiv preprint arXiv:2504.05408.
129. Raheja, T., Pochhi, N., & Curie, F. (2024). Recent advancements in LLM red-teaming: Techniques, defenses, and ethical considerations. arXiv preprint arXiv:2410.09097.
130. Raman, D., Madkour, N., Murphy, E., Jackson, K., et al. (2025). Intolerable risk threshold recommendations for artificial intelligence. arXiv preprint arXiv:2503.05812.
131. Reid, A., O'Callaghan, S., Carroll, L., & Caetano, T. (2025). Risk analysis techniques for governed LLM-based multi-agent systems. arXiv preprint arXiv:2508.05687.
132. Riyadh, R. (2025). Cognitive honeypots AI-enhanced deception for proactive threat hunting. *AlKadhim Journal for Computer Science*.
133. Roy, S. (2025). Persuasiveness and bias in LLM: Investigating the impact of persuasiveness and reinforcement of bias in language models. arXiv preprint arXiv:2508.15798.

134. Sabour, S., Liu, J., Liu, S., Yao, C., Cui, S., et al. (2025). *Human decision-making is susceptible to ai-driven manipulation*. arXiv preprint arXiv:2502.07663.

135. Sachdeva, A., Saravanan, R., Sarkar, G., Vemuri, K., et al. (2025). BEACON: A unified behavioral-tactical framework for explainable cybercrime analysis with large language models. arXiv preprint arXiv:2512.06555.

136. Sartori, G., & Orrù, G. (2023). Language models and psychological sciences. *Frontiers in Psychology*.

137. Schmitt, M., & Flechais, I. (2024). Digital deception: Generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*.

138. Schneider, J., Haag, S., & Kruse, L. (2024). Negotiating with LLMs: Prompt hacks, skill gaps, and reasoning deficits. *International Conference on Computer Supported Cooperative Work and Social Computing*.

139. Schoenegger, P., Salvi, F., Liu, J., Nan, X., et al. (2025). *Large language models are more persuasive than incentivized human persuaders*. arXiv preprint arXiv:2505.09662.

140. Shen, G., Cheng, S., Zhang, K., Tao, G., An, S., Yan, L., et al. (2024). *Rapid optimization for jailbreaking LLMs via subconscious exploitation and echopraxia*. arXiv preprint arXiv:2402.05467.

141. Shen, Z., Yan, S., Zhang, Y., Luo, X., Ngai, G., et al. (2025). "it warned me just at the right moment": Exploring LLM-based real-time detection of phone scams. *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*.

142. Shim, H., Park, H., Lee, K., Park, J., & Kang, S. (2024). A persuasion-based prompt learning approach to improve smishing detection through data augmentation. arXiv preprint arXiv:2411.02403.

143. Siemerink, A., Jansen, S., & Labunets, K. (2024). The dual-edged sword of large language models in phishing. *Nordic Conference on Secure IT Systems*.

144. Singh, C., Inala, J., Galley, M., Caruana, R., et al. (2024). *Rethinking interpretability in the era of large language models*. arXiv preprint arXiv:2402.01761.

145. Singh, G., Singh, P., & Singh, M. (2025). *Advanced real-time fraud detection using RAG-based LLMs*. arXiv preprint arXiv:2501.15290.

146. Singh, S., Abri, F., & Namin, A. (2023). Exploiting large language models (LLMs) through deception techniques and persuasion principles. *2023 IEEE International Conference on Big Data*.

147. Sison, A., Daza, M., Gozalo-Brizuela, R., et al. (2024). ChatGPT: More than a “weapon of mass deception” ethical challenges and responses from the human-centered artificial intelligence (HCAI) perspective. *Journal of Human - Computer Interaction*.

148. Starace, J., & Soule, T. (2026). *Intentional deception as controllable capability in LLM agents*. arXiv preprint arXiv:2603.07848.

149. Street, W. (2024). LLM theory of mind and alignment: Opportunities and risks. arXiv preprint arXiv:2405.08154.

150. Sun, G., Zhan, X., & Such, J. (2024). Building better ai agents: A provocation on the utilisation of persona in LLM-based conversational agents. *Unable to Determine the Complete Publication Venue*.

151. Sun, Y., & Wang, T. (2025). *Be friendly, not friends: How LLM sycophancy shapes user trust*. arXiv preprint arXiv:2502.10844.

152. Tan, X., See, K., & Kok, S. (2024). ScamGPT-j: Inside the scammer’s mind, a generative AI-based approach toward combating messaging scams. arXiv preprint arXiv:2412.13528.

153. Tang, Y., Wang, Y., Qiu, L., Gao, W., Ma, Y., Chen, B., et al. (2026). *VirtualCrime: Evaluating criminal potential of large language models via sandbox simulation*. arXiv preprint arXiv:2601.13981.

154. Tarsney, C. (2025). Deception and manipulation in generative AI: C. tarsney. *Philosophical Studies*.

155. Timm, J., Talele, C., & Haimes, J. (2025). Tailored truths: Optimizing LLM persuasion with personalization and fabricated statistics. arXiv preprint arXiv:2501.17273.

156. Trinh, Q., Zarin, S., & Rezapour, R. (2025). Master of deceit: Comparative analysis of human and machine-generated deceptive text. *Proceedings of the 17th ACM Web Science Conference*.

157. Tshimula, J., Nkashama, D., Muabila, J., et al. (2024). Psychological profiling in cybersecurity: A look at LLMs and psycholinguistic features. *Unable to Determine the Complete Publication Venue*.

158. Ugarte, M., Valle, P., Parejo, J., Segura, S., et al. (2025). Astral: A tool for the automated safety testing of large language models. *Proceedings of the 34th Annual ACM Symposium on Applied Computing*.

159. Wang, L., Ma, Y., Gao, R., Guo, B., Zhu, H., Fan, W., et al. (2024). *Megafake: A theory-driven dataset of fake news generated by large language models*. arXiv preprint arXiv:2408.11871.
160. Wang, N., Walter, K., Gao, Y., & Abuadbbba, A. (2025). *Large language model adversarial landscape through the lens of attack objectives*. arXiv preprint arXiv:2502.02960.
161. Wang, R., Chen, M., Chang, K., & Hsueh, T. (6151). SpAI phishing: A design and evaluation of AI-assisted large spear phishing cybersecurity drill. *Available at SSRN 6151594*.
162. Wang, S., Zhao, Y., Liu, Z., Zou, Q., & Wang, H. (2025). *Sok: Understanding vulnerabilities in the large language model supply chain*. arXiv preprint arXiv:2502.12497.
163. Wang, X., Zhang, W., Koneru, S., Guo, H., et al. (2025). Have LLMs reopened the pandora's box of AI-generated fake news? *Proceedings of the North American Chapter of the Association for Computational Linguistics*.
164. Wang, Z., Cao, Y., & Liu, P. (2024). Hidden you malicious goal into benign narratives: Jailbreak large language models through logic chain injection. arXiv preprint arXiv:2404.04849.
165. Wang, Z., Xie, W., Wang, B., Wang, E., Gui, Z., Ma, S., et al. (2024). *Foot in the door: Understanding large language model jailbreaking via cognitive psychology*. arXiv preprint arXiv:2402.15690.
166. Williams, M., Carroll, M., Narang, A., Weisser, C., et al. (2024). *On targeted manipulation and deception when optimizing LLMs for user feedback*. arXiv preprint arXiv:2411.02306.
167. Winter, J. de, Hancock, P., & Eisma, Y. (2025). ChatGPT and academic work: New psychological phenomena. *AI & SOCIETY*.
168. Wu, X., Hong, G., Chen, P., Chen, Y., Pan, X., et al. (2025). *Prison: Unmasking the criminal potential of large language models*. arXiv preprint arXiv:2506.16150.
169. Wu, Y., Pan, X., Hong, G., & Yang, M. (2025). Opendedeception: Benchmarking and investigating ai deceptive behaviors via open-ended interaction simulation. arXiv preprint arXiv:2504.13707.
170. Xu, Z., Yu, C., Fang, F., Wang, Y., & Wu, Y. (2023). *Language agents with reinforcement learning for strategic play in the werewolf game*. arXiv preprint arXiv:2310.18940.
171. Xue, Y., Wang, J., Yin, Z., Ma, Y., Qin, H., Tao, R., et al. (2025). Dual intention escape: Penetrating and toxic jailbreak attack against large language models. *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*.
172. Yang, X., Zhou, B., Tang, X., Han, J., & Hu, S. (2026). Exploiting synergistic cognitive biases to bypass safety in LLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*.
173. Yu, J., Yu, Y., Wang, X., Lin, Y., Yang, M., Qiao, Y., et al. (2024). *The shadow of fraud: The emerging danger of ai-powered social engineering and its possible cure*. arXiv preprint arXiv:2407.15912.
174. Yu, S., Carroll, F., & Bentley, B. (2024). Trust and risk: Psybersecurity in the AI era. *Psybersecurity*.
175. Zard, L. (2025). Online exploitation: Drawing the line for surveillance advertising in europe. *Journal of Advertising*.
176. Zeng, Y., Lin, H., Zhang, J., Yang, D., Jia, R., et al. (2024). How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing LLMs. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.
177. Zhan, X., Xu, Y., Abdi, N., Collenette, J., & Sarkadi, S. (2025). Banal deception and human-AI ecosystems: A study of people's perceptions of LLM-generated deceptive behaviour. *Journal of Artificial Intelligence Research*.
178. Zhang, H., Zhu, C., Wang, X., Zhou, Z., Yin, C., Li, M., et al. (2024). *Badrobot: Jailbreaking embodied LLMs in the physical world*. arXiv preprint arXiv:2407.20242.
179. Zhang, K., Su, Z., Chen, P., Bertino, E., Zhang, X., et al. (2025). *LLM agents should employ security principles*. arXiv preprint arXiv:2505.24019.
180. Zhang, L., Wang, H., Cheng, L., Deng, L., & Ward, T. (2025). *Adversarial testing in LLMs: Insights into decision-making vulnerabilities*. arXiv preprint arXiv:2505.13195.
181. Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H., et al. (2025). The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. *Proceedings of*.
182. Zhang, W., Sheng, Q., Alhazmi, A., & Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology*.

183. Zhao, X., Yang, X., Pang, T., Du, C., Li, L., Wang, Y., et al. (2024). *Weak-to-strong jailbreaking on large language models*. arXiv preprint arXiv:2401.17256.
184. Zhou, Y., Bao, H., Huang, Y., Guo, K., Liang, Z., et al. (2025). Emergent deceptive behaviors in reward-optimizing LLMs. *Unable to Determine the Complete Publication Venue*.
185. Zucca, M., & Fiorinelli, G. (2025). Regulating AI as a cybersecurity defense: Fighting the misuse of generative AI for cyber attacks and cybercrime. *Technology and Regulation*.
186. Zucca, M., & Gaia, F. (2025). Regulating AI to combat tech-crimes: Fighting the misuse of generative AI for cyber attacks and digital offenses. *Technology and Regulation*.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.