CogSLLaM: Cognitive Security for Large Language Models

Janine Mator, Chelsea Johnson, Vasanth Sarathy, Erin Roper, Antonio Piazza, Ethan Irby and Kimberly Ferguson-Walter

Leidos
kimberly.j.fergusonwalter@leidos.com

Abstract

Large Language Models (LLMs) are increasingly employed across general purpose, expert, and domain-intensive user contexts; in all instances, users rely on these systems for information summarization, task planning, and problem-solving, and even for advice and recommendations related to decision-making. LLMs are becoming embedded in highly technical domains, supporting cyber security, code development, health care, organizational policy analysis, and the interpretation, specification, and implementation of complex technical requirements. This expanding range of applications underscores the growing role of LLMs as general-purpose cognitive and technical support tools. However, increasing reliance on LLMs also introduces cognitive security risks, as erroneous or misleading output can shape user understanding, judgment, and decision-making in ways that produce real-world consequences. In this paper, we examine cognitive security challenges in LLM applications, their downstream effects on users, relevant technical approaches for addressing these risks, representative use cases, and our proposed Cognitive Security for LLMs framework. This framework focuses on three dimensions along which cognitive security may be threatened: informational, semantic, and stylistic. We conclude with key takeaways and future directions for reducing misleading and potentially harmful chatbot interactions across everyday and high-stakes contexts.

Introduction & Motivation

Large Language Models (LLMs) can potentially manipulate human cognition and sensemaking, jeopardizing individual and collective reasoning about the world. This may occur as the result of deliberate use of adversarial methods (e.g., noise against model data), purposeful model construction (e.g., deceptive design) or inadvertent consequences (e.g., model training/data).

Semantic decisions, such as word choice and ordering, can change the interpretation of information by using persuasive, emotionally triggering, or loaded terms. Furthermore, LLM stylistic features modulate the cognitive status of a statement. For example, the same propositional content, presented in differentiated typographic styles, will be processed differently based on perceived credibility, memorability, emotional impact, and actionability.

Whether deliberate or unintentional, these trends represent a risk to cognitive security and may damage our capacity to make decisions that are not unduly influenced by technologies designed to help us. With the rapid adoption of LLMs in various capacities, understanding these potential impacts is critical to ensuring decision-making autonomy.

We argue that research in human cognition provides a critical bridge between cognitive security and cybersecurity practice by offering tangible, empirically sound methods for safeguarding human decision-making from distortion during information processing. Application of work on sensemaking, persuasion, deception, and human factors with the computational capabilities of Artificial Intelligence (AI) provides a foundation for advancing cognitive security during this period of rapid technological development and adoption.

In this work, we have defined three main dimensions along which cognitive security is threatened by LLMs—informational, semantic, and stylistic. We have also outlined a technical solution to mitigate these risks and preserve user decision-making autonomy.

By mitigating the influence of inaccurate or misleading interactions with LLMs, cognitive security safeguards can support robust sensemaking—the process through which individuals derive meaning from complex, ambiguous, or unexpected information (Weick 1995). In this context, sensemaking refers specifically to the process through which users derive insights from LLM-generated outputs and subsequently incorporate those insights into decision-making. Because the content, semantics, and stylistic features of information shape how users process that information and form related judgments, LLM-based chatbots become participants in the user's sensemaking process and, consequently, potential sources of risk to cognitive security. Accordingly, it is essential to identify the origins and categories of threats to sensemaking in LLM interactions, assess their impacts on users, and evaluate the State of the Art (SOTA) in maintaining cognitive security during these exchanges.

Threats to Cognitive Security

Threats to cognitive security in LLM interactions originate in the design and implementation of the models themselves. Recent studies have identified several limitations associated with agentic backends, including priming effects (Shaki, Kraus, and Wooldridge 2023), reduced summarization accuracy for long documents (Tang et al. 2023), overreliance on lexical similarity in source materials (Luo, Xie, and Ananiadou 2023), and the justification of false statements (Danry et al. 2025). Concerns about hallucinations have prompted the development of numerous standalone fact-checking tools (e.g., OpenFactCheck (Iqbal et al. 2024); FactCheck-GPT (Wang et al. 2024); FacTool (Chern et al. 2023)). However, such tools are insufficient for supporting sensemaking of complex subject matter, as they cannot address users' follow-up questions and often require the use of a separate application that disrupts the user's workflow. Furthermore, although mission statements from major AI companies such as OpenAI express a commitment to improving LLM factuality by reducing hallucination rates, these safeguards do not necessarily address how accurately information is interpreted. Factual information may be over-simplified to the point of misinterpretation. Additionally, well-documented attributes of LLM-based agents, such as sycophancy and overconfidence, can contribute to faulty interpretations of otherwise verified claims (Danry et al. 2025).

Additional threats to cognitive security stem from limitations of the user, including attentional resource constraints, heuristic thinking, and cognitive biases. Lower attentional control and working memory capacity have been linked to poorer learning outcomes and reduced active participation in the learning process (Choi, Syeda Sabrina Akter, and Anastasopoulos 2024; Diamond 2013), suggesting similar challenges when processing LLM-generated output. Individuals may also employ heuristic techniques—mental shortcuts or "rules of thumb"—to reduce cognitive load and expedite decision-making; however, such strategies can promote shallow reasoning and reinforce cognitive biases (Tversky and Kahneman 1974). Prevalent examples include anchoring bias, in which individuals over-rely on initial information at the expense of accounting for later information, and confirmation bias, in which individuals seek and retain information that supports existing beliefs (Pohl 2004). Automation bias, the tendency to over-trust and over-rely on computer-generated information, is another form of cognitive bias that is particularly relevant to LLM chatbot users (Buçinca, Malaya, and Gajos 2021).

Additional challenges to cognitive security arise from User Interface (UI) designs that direct attentional focus toward arbitrary, unimportant, or unhelpful elements of model output without sufficient context or evaluation of the underlying claims. When interfaces misdirect users' limited attentional resources (e.g., by providing claims in bullets without indicators of uncertainty or alternative viewpoints), users may be more likely to accept LLM outputs without sufficient critical evaluation.

Impact on LLM Users

Negative outcomes of LLM chatbots on the user often stem from a trade-off between efficient sensemaking and cognitive security. The more rapidly users seek to process information from LLMs, the fewer opportunities there are to engage in independent critical thinking and to challenge the depth of understanding.

Poor decision-making is one consequence of threats to cognitive security during LLM interactions. For example, based on an LLM's generalized summary of research methods within a broad domain, a user may select a method with less generalizability than alternatives they might have identified through more robust reasoning. Additionally, the more users offset the burden of cognitive workload through LLMs (i.e., cognitive offloading), the more they may reinforce cognitive biases and heuristic techniques that promote flawed sensemaking. In this way, cognitive offloading and automation bias can become a vicious cycle that worsens over time (Gerlich 2025). The combination of short- and long-term effects on LLM users highlights the need for appropriate, accessible safeguards for cognitive security, as well as Human Subjects Research (HSR) to test and compare potential solutions.

Related Solutions

Current solutions to protect cognitive security during LLM interactions leave room for improvement. Beyond the fact-checking tools mentioned previously, Reinforcement Learning with Human Feedback (RLHF) involves learning from human preferences through a reward model trained with human-rated outputs and has been suggested as a best practice. Although one camp embraces and adopts RLHF (Ouyang et al. 2022), another suggests that RLHF is conceptually and practically inadequate for capturing the complexity of human ethics and for delivering robust AI safety (Dahlgren Lindström et al. 2025).

Recent years have also seen a handful of proposed tools as user-facing solutions to support better sensemaking with LLMs. Sensecape, seeks to aid users' workflow by organizing information into multiple levels of abstraction (Suh et al. 2023). This includes a hierarchy view for more high-level understandings of concepts within a domain, as well as a canvas view, which functions as a kind of whiteboard for users to conduct more granular diagramming. In their evaluation, users explored more domain-specific terms and demonstrated deeper levels of understanding with Sensecape compared to a conversational chatbot interface alone (Suh et al. 2023). Similarly, ScholarMate (Jiang et al. 2025) is an interface designed for users to explore working theories by interacting with snippets of text from source documents on a non-linear interface. PaperWeaver (Lee et al. 2024), another proposed sensemaking solution, was designed to notify users of relevant research papers and summarize how each one relates to the user's existing library.

Such tools are admirable in their aim to promote low- and high-level conceptualization and user engagement beyond that of passive chatbot interactions. However, the focus of these tools remains on aiding sensemaking through

virtual workspaces rather than protecting cognitive security outright, and the sophistication of some features may be off-putting to novice users. Our proposed approach, Cognitive Security for Large Language Models (CogSLLaM) addresses these gaps in current methodologies and sensemaking tools by combining claim validation, cognitive and behavioral science, and an interactive UI with default and advanced feature modes for accessibility to all users.

Challenge of Multi-document Summarization

LLMs are being increasingly used for various tasks that involve processing large bodies of text and providing summaries and answers about these texts. For example, LLMs are being used to read resumes, identify candidates best suited for a job, summarize candidates' relevant experiences, and answer questions about candidates with certain skills or experiences. Similarly, summarization tasks are popular in medical, legal and defense/intelligence contexts for reducing the human burden of reviewing documents at scale.

However, there is an extensive list of known problems with LLM summarization tools: 1) hallucinations that fabricate citations, invent new information, and draw incorrect causal inferences (Belém et al. 2025; Liu et al. 2026); 2) lack of faithfulness to sources by changing the meaning of source claims, misinterpreting relationships between entities, and providing incorrect attributions (Tam et al. 2023); 3) overgeneralizing and oversimplifying by making generalized claims beyond sources, omitting key caveats and limitations, and flattening nuanced arguments (Peters and Chin-Yee 2025); 4) omitting critical information by missing numerical values, ignoring methodological details, and excluding minority viewpoints (Asgari et al. 2025); 5) lack of uncertainty awareness (e.g., overconfidence, susceptibility to distractors, and lack of calibration; (Chhikara 2025)); 6) limitations of their context window resulting in a loss of global structure and inability to maintain cross-section dependencies (Sonowal and Sadhu 2025); and
7) data biases that emphasize or deemphasize certain viewpoints or alter interpretations toward certain political or social frames (Vijay, Priyanshu, and KhudaBukhsh 2025). For these reasons, our approach focuses on addressing the challenges of multi-document summarization as an initial use case.

Cybersecurity: A High Impact Use Case

Working in cybersecurity imposes stress from high stakes, time pressure, and high cognitive load, making effective decision-making a paramount concern. In cybersecurity applications, LLMs are increasingly used to summarize and document code and logs, triage issues, and generally serve as a Human-AI interface to highly sophisticated cybersecurity tooling (e.g., fuzzing software); therefore, technical solutions to improve document summarization while preserving cognitive security are a critical challenge in cyber. Below, we provide examples of specific representative use cases in cyber through the lens of multi-document summarization:

- *Threat intelligence analysis* requires synthesizing heterogeneous sources, including incident reports, malware analyses, telemetry, and geopolitical context. Here, LLMs are used to generate analytic summaries.
- *Vulnerability triage* requires prioritizing remediation across large volumes of advisories. Here, LLMs are used to summarize Common Vulnerabilities and Exposures (CVEs), interpret exploit conditions, and recommend prioritization.
- *Incident response* involves high-volume data combined with ambiguous alerts. LLMs are being explored to assist with log interpretation and anomaly explanation.

Cross-Use-Case Implications

Across threat intelligence, vulnerability management, incident response, and document summarization tasks, many of the most consequential failures arise from calibration errors rather than blatant hallucinations. Studies have demonstrated that specific cognitive vulnerabilities, including loss aversion and confirmation bias, are especially relevant to cyber tasks (Ferguson-Walter, Major., and Roper 2026; Gutzwiller et al. 2024).

CogSLLaM Approach

Threat Model

We conceptualize the LLM as a potential threat actor capable of generating cognitive security risks, whether intentionally or unintentionally. To systematically assess and calibrate LLM output, we introduce a Cognitive Security framework composed of three dimensions: informational, semantic, and stylistic. Beyond detecting and mitigating cognitive security vulnerabilities, the framework also provides users with explainability and interpretability regarding how these interventions are performed. As shown in Figure 1, the CogSLLaM architecture operationalizes these dimensions in parallel.

Cognitive Security Framework

We identify three key dimensions along which an LLM output can mislead or influence a human user:

1. **Informational:** The information provided by the LLM is inaccurate, flawed, or fabricated.
2. **Semantic:** The output is presented in a way that misconstrues or shapes the meaning of the information, impacting the user's interpretation.
3. **Stylistic:** The presentation or format places arbitrary emphasis on word components that misrepresent authority, significance, or credibility.

Informational Analysis Our thesis is that LLM outputs for summarization tasks (and potentially a range of other interpretive tasks) can be viewed through the lens of argumentation. That is, when summarizing a set of source documents, the LLM makes claims about a topic (ostensibly sourced from given source documents) and provides reasons for why one should accept those claims. Argumentation is a core social reasoning skill possessed by humans and present in nearly every task-based conversation. We therefore model the informational accuracy of LLM output (e.g., summaries)

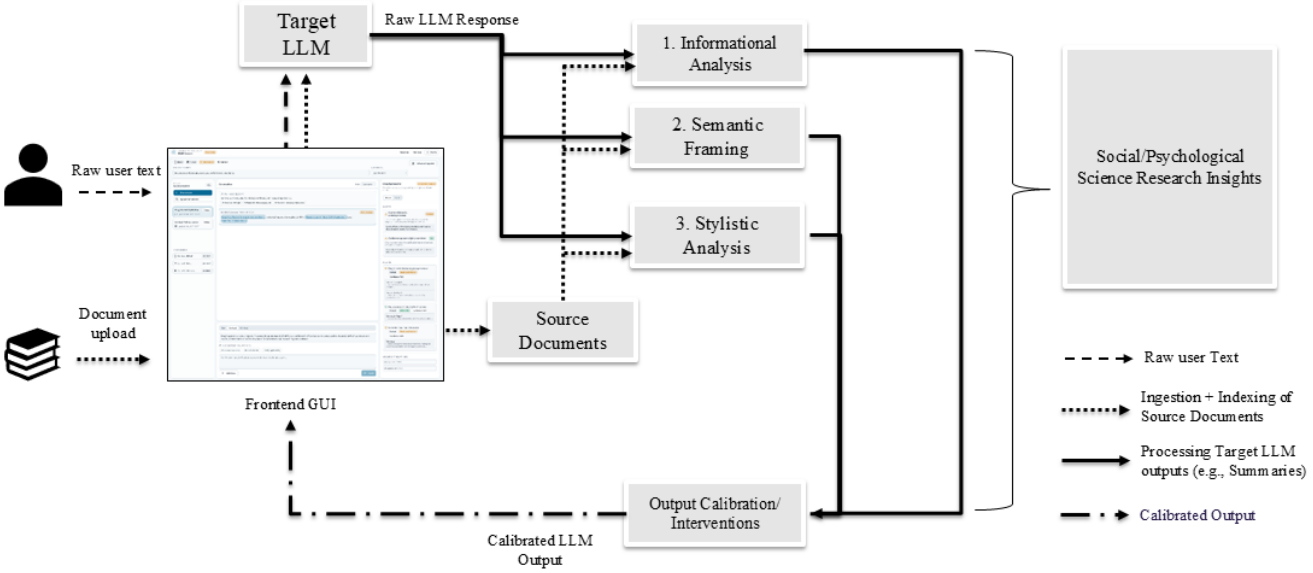

Figure 1: CogSLLaM System Architecture

based on the nature and quality of the argumentative elements they possess. Just as CogSLLaM extracts claims from source documents, it also extracts claims from LLM-generated output. Given the source documents, CogSLLaM then quantifies the strength of the summary based on the thoroughness of its argument (i.e., the presence of incorrect information, overly broad claims, and omission of important counterclaims).

To enable automated argumentation as part of the informational analysis, we will perform the task of argument mining—extracting argument premises, conclusions, and semantic and structural relations between these components from a given natural language input (Sarathy et al. 2022; Lawrence and Reed 2019). The task of argument mining involves: (1) identifying argument discourse units (ADUs) as spans of text such that each span minimally captures a single argument (Stede et al. 2016); (2) identifying relationships among ADUs; and (3) classifying both the type of ADU (e.g., claim type) and the type of relation (e.g., support/attack). We adopt a one-shot, end-to-end approach in which an argumentation model, given an argumentative natural language input, performs all three tasks simultaneously–separating ADUs from the main text, identifying their relationships, and classifying each relation. The output of the argument mining process can be represented as a graph with nodes representing ADUs (claims) and edges representing inferential or evidential relations between claims.

CogSLLaM performs argument mining on both the summaries generated by the target LLM and the source documents. By focusing on the claims alone (ADUs), this approach enables evaluation of the accuracy of the claims in the raw LLM-generated response summaries by comparing them to those in the source documents, similar to approaches such as VeriFact (Liu et al. 2025). Unlike VeriFact, however, our approach works with claims and their relationships, providing both fact and argument-level verification. This allows assessment of whether the argument presented in the summary is faithful to those in the source documents by comparing their respective argument graphs. Moreover, argument graphs support graph-based reasoning techniques for detecting argument flaws, contradictions, and other cognitive security-relevant features. For example, circular arguments are simply cycles in the graph, while strawman arguments are when a node B attacks A', but A' is a weaker version of A in the source.

For example, consider an LLM that summarizes a set of intelligence source documents as follows: "*Recent intelligence reports indicate that the cyber-espionage group, Orchid Spider, conducted a widespread intrusion campaign targeting U.S. energy infrastructure in early 2025.*" In this illustrative example, CogSLLaM will be able to extract similar claims and compare them against those made directly in the source documents. Thus, if there is only evidence that two regional utilities experienced unauthorized access, then we will be able to flag the claim as overgeneralized and not supportive of a claim of "widespread intrusion." Moreover, facts in source documents may both support and attack contradictory elements of the LLM summary. Our computational argumentation approach mines for claims and support/attack relationships and performs reasoning over these representations to provide corrections such as identifying when a claim omits critical counter evidence and when confidence should be decreased.

Semantic Analysis LLMs increasingly mediate how people encounter explanations, recommendations, and assertions. This mediation is not only informational (the "what" e.g., the arguments, claims, explanations, or instructions the model provides) but also through semantics, the overall

meaning of text as a function of its parts. These semantics affect cognitive security by shaping decision-making, belief formation and eventually action, often without users' awareness or consent (Singh and Namin 2025).

For example, whether in a benign or malicious context, persuasion principles can be used to manipulate cognitive and decision-making processes (Cialdini 1995). LLM output can sound like neutral assistance while quietly shaping user perception of what is normal, credible, urgent, or morally consistent (Muhanad et al. 2025). The following output demonstrates persuasion principles in recommending a security monitoring plan: "*Most people in your position choose the premium monitoring plan because it is the safest option. Security experts generally recommend acting early because delaying could leave you exposed.*" In this notional example, social validation ("most people"), authority ("security experts"), and scarcity ("delaying could leave you exposed") lead to a prespecified conclusion. Similarly, polarizing terms, emotional triggers, and other surface level details of information presentation can have undue influence over how findings are interpreted.

CogSLLaM addresses these issues by extracting arguments in the source documents, identifying meaningful semantic differences via internal Natural Language Processing (NLP) modules, and providing feedback and information through the UI.

Stylistic Analysis Style (i.e., typography) is the structure and arrangement of visual language that provide design markers and shapes how readers approach and understand text (Baines and Haslam 2005). Visual cues like different fonts, sizes, styles, colors, and layout can signal whether to skim quickly or read more carefully and thoughtfully (Folarin and Okonkwo 2025). Particularly when extracting output from a LLM UI (e.g., for a Word document) content may be automatically stylized with bullets, emphases (e.g., bold, italics) and other stylistic elements. These salient features draw users' attention and lead to attribution of authority, priority, or significance where none might exist. While the intent of style is to provide "shortcuts", thus reducing cognitive effort, arbitrary application can misleadingly influence users and impact user decision-making.

A library of basic prompts included to direct the LLM as to a normalized output style is one initial step in calibrating stylistic variance.

Behavioral Interventions and Cognitive Cues

Figure 2 illustrates our framework for applying cognitive science principles to the development of LLMs that promote cognitive security for users. Specifically, we emphasize the utility of technical mechanisms that engage critical thinking processes to address informational challenges. For example, to address potential misinformation resulting from sourcing information, our solution performs behavioral intervention through the calibrated output specifying the nature of the sources, contextualizing their claims, and alerting users to the full text or original location. Semantic challenges are mitigated through cognitive cues that encourage meaning-based examination of content by providing context for interpretation, moderating confidence, and identifying connotationally loaded terminology, for example. Finally, stylistic challenges are managed by imposing standardized presentation and format to ensure that no subset of the information is given undue consideration by the user.

CogSLLaM begins to mitigate many of these risks by decomposing summaries into explicit claims, mapping them to supporting and opposing evidence, and highlighting elements critical to maintaining cognitive security. This includes surfacing counterclaims, introducing calibrated uncertainty markers to help preserve analytic reasoning, distinguishing model inference from source-backed facts, and enforcing explicit linkage between generated claims and source information, featuring key contextual information such as uncertainty, and presenting competing hypotheses when appropriate. Cognitive security requires analysis of framing and presentation alongside accuracy of the information. Some of these interventions (e.g., changing/deleting incorrect information) are addressed through calibration of the LLM output, meaning that it is automatically corrected before being presented to the user. However, some interventions (e.g., highlighting polarizing terms, or persuasive tactics) instead require a UI solution that provides explicit indicators to the user regarding potential cognitive security risks. Several factors, including context, user characteristics (e.g., role, expertise), task difficulty (Eigner and Händler 2024), and user goals, influence whether a mitigation should be applied automatically or presented as a user-selectable option. Ensuring these algorithmic and interface interventions are backed by existing cognitive science findings and examining the impact on decision-making through user studies are the critical next steps.

Conclusions & Future Work

The Cognitive Security Institute is a non-profit organization created to highlight these issues. They define the problem space as follows: *"In the digital age, our thoughts, beliefs, and decisions are under constant attack. Sophisticated adversaries use cognitive warfare to erode trust, manipulate perceptions, and destabilize societies. These threats target past our systems to our very minds. Traditional cybersecurity alone isn't enough We need cognitive security"(CSI 2025).*

We argue that the need for cognitive security expands beyond sophisticated cognitive warfare—that AI usage and AI algorithms themselves are also subtly influencing human decision-making outcomes in ways not yet fully understood. This is a much needed and critical area of research. Advancing this area will require coordinated efforts across the research community to build technology that helps preserve the agency of the decision-maker and provide decision advantage.

Human-AI teaming is increasingly pervasive as an area of concern to the general public. In this paper, we have focused on a subset of implications of Human-AI teaming—namely the Cognitive Security impacts of LLM usage—and examined the (sometimes seemingly invisible) ramifications to human cognition and decision-making. As Human-AI teams increase and co-mingle with human teams, and AI agent

Human Decision-Making Challenges with LLM Usage	Foundational Behavioral Science	Technical Mechanism for Behavioral Intervention	Ensure Cognitive Security
Informational Analysis (Overreliance on System 1 Automaticity)	System 2 to Engage Cricital Thinking	• Source Information • Evidence for Claims • Inconsistencies • Factual Inaccuracy • Algorithmic Biases • Human Cognitive Limitations • Relationships	• Characterize sources that over/under claim • Identify/correct hallucinations • Highlight omission of critical information • Correct flawed linkages
Semantic Analysis (Focusing interpretation on the most fundamental points)	Meaning-Based Content Examination	• Connotation • Emotional Intensity/Affective Loading • Coherence • Context • Persuasion • Sycophancy • Narrative Structure	• Imposes missing context for interpretation • Moderate overconfidence • Highlight potentially polarizing or persuasive terms
Stylistic Analysis (Paying attention to presentation details)	Attention Allocation	• Order • Emphasis (e.g., bolding to draw attention) • Consistency • Presentation	• Standardized presentation and format

Figure 2: Application of Cognitive Science to Address Cognitive Security Challenges of LLM Usage

teams, these recursive patterns of interactions will continue to create new, and exponentially complicated challenges, not yet identified. AI algorithms are not typically designed with cognitive security in mind, and there have been many instances in which LLMs have produced disinformation and misinformation, sometimes with catastrophic outcomes.

CogSLLaM is an initial attempt to focus on the need for both algorithmic and UI solutions that explicitly address cognitive security to: (1) identify the issues, (2) understand what this means for Cognitive Security, (3) modify/improve our tech solutions based on behavioral science, and (4) establish and measure effectiveness of solutions. Future work includes further investigation of remediations to LLM shortfalls and limitations and their impact on human decision-making. Future work will further investigate what interventions should be automated, and what should be based on user choice, and the impact of explainability and traceability on both novice and expert users. Likely the system will require an "easy mode" for those who desire less information and less interaction with systems and an "expert mode" for those who want to see and interact with each decision and change made by the system. Next steps for this research include detailing the technical interventions needed to automatically calibrate and normalize LLM output, as well as interactive UI overlays that caveat and alert users to important aspects of the output. Human subjects research is planned to test and improve these interventions, with the ultimate goal of preserving human decision autonomy and advancing decision advantage.

References

Asgari, E.; Montaña-Brown, N.; Dubois, M.; Khalil, S.; Balloch, J.; Yeung, J. A.; and Pi-menta, D. 2025. A Framework to Assess Clinical Safety and Hal-lucination Rates of LLMs for Medical Text Summarisation. *NPJ Digital Medicine*, 8(1): 274.

Baines, P.; and Haslam, A. 2005. *Type & typography*. Laurence King Publishing.

Belém, C. G.; Pezeshkpour, P.; Iso, H.; Maekawa, S.; Bhutani, N.; and Hruschka, E. 2025. From Single to Multi: How LLMs Hallucinate in Multi-Document Summarization. In *Findings of the Association for Computa-tional Linguistics: NAACL 2025*, 5291–5324. Albuquerque, New Mexico: Association for Computational Linguistics.

Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–21.

Chern, I.-C.; Chern, S.; Chen, S.; Yuan, W.; Feng, K.; Zhou, C.; He, J.; Neubig, G.; and Liu, P. 2023. FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. ArXiv preprint. arXiv:2307.13528.

Chhikara, P. 2025. Mind the Confidence Gap: Overconfidence, Calibration, and Distractor Effects in Large Language Models. ArXiv preprint. arXiv:2502.11028.

Choi, A. S.; Syeda Sabrina Akter, J.; and Anastasopoulos, A.-t. 2024. The LLM Effect: Are Humans Truly Using LLMs, or Are They Being Influenced by Them In-stead? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 22032–22054. Miami, FL: Association for Computational Linguistics.

Cialdini, R. B. 1995. Principles and Techniques of Social Influence. In Tesser, A.-h.; and Tesser, A., eds., *Advanced Social Psychology*, 257–281. Mahwah, NJ: ErlbaumCialdini. McGraw-Hill. (Wiley Online Library).

CSI. 2025. Cognitive Security Institute. Https://www.cognitivesecurityinstitute.org/.

Dahlgren Lindström, A.; Methnani, L.; Krause, L.; Ericson, P.; de Rituerto de Troya, I. n. M.; Coelho Mollo, D.; and Dobbe, R. 2025. Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through Reinforcement Learning from Human Feedback: Helpful, harmless, honest? Sociotechnical limits of AI alignment... *Ethics and Inf. Technol.*, 27(2).

Danry, V.; Pataranutaporn, P.; Epstein, Z.; Groh, M.; and Maes, P. 2025. Deceptive Explanations by Large Language Models Lead People to Change Their Beliefs About Misinformation More Often than Honest Explanations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–31. New York: Association for Computing Machinery.

Diamond, A. 2013. Executive Functions. *Annual Review of Psychology*, 64(1): 135–168.

Eigner, E.; and Händler, T. 2024. Determinants of LLM-Assisted Decision-Making. ArXiv preprint. arXiv:2402.17385.

Ferguson-Walter, K. J.; Major., M. M.; and Roper, E. R. 2026. A Cyber+Human Data Windfall to Revolutionize Cyberpsychology, Predictive Modeling, and Cyber Defense. *Computational Brain and Behavior. Forthcoming*.

Folarin, A. O.; and Okonkwo, C. N. 2025. How Do Typography and Layout Signal Identity and Genre and Anticipate Reading Strategies? *Open Journal of In-formatics and Technology*, 2(4): 1–10.

Gerlich, M. 2025. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, 15(1): 6.

Gutzwiller, R. S.; Rheem, H.; Ferguson-Walter, K. J.; Lewis, C. M.; Johnson, C. K.; and Major, M. M. 2024. Exploratory Analysis of Decision-Making Biases of Professional Red Teamers in a Cyber-Attack Da-taset. *Journal of Cognitive Engineering and Decision Making*, 18(1): 37–51.

Iqbal, H.; Wang, Y.; Wang, M.; Georgiev, G. N.; Geng, J.; Gurevych, I.; and Nakov, P. 2024. OpenFactCheck: A Unified Framework for Factuality Evaluation of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 219–229. Miami, FL: Association for Compu-tational Linguistics.

Jiang, D. L.; Ye, S.; Zhao, L.; and Gu, B. 2025. Do Reductions in Search Costs for Partial Information on Online Platforms Lead to Better Consumer Decisions? Evi-dence of Cognitive Miser Behavior from a Natural Experi-ment. *Information Systems Research*, 36(3): 1780–1798.

Lawrence, J.; and Reed, C. 2019. Argument Mining: A Survey. *Computational Linguistics*, 45(4): 765–818.

Lee, Y.; Kang, H. B.; Latzke, M.; Kim, J.; Bragg, J.; Chang, J. C.; and Siangliulue, P. 2024. PaperWeaver: Enriching Topical Paper Alerts by Contextualizing Recommended Papers with User-collected Papers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–19. ACM.

Liu, S.; Gao, Y.; Li, S.; Wang, P.; and Wang, T. 2026. A Hallucination Detection and Mitigation Framework for Faithful Text Summarization Using LLMs. *Sci-entific Reports*, 16: 1374.

Liu, X.; Zhang, L.; Munir, S.; Gu, Y.; and Wang, L. 2025. VeriFact: Enhancing Long-Form Factuality Evaluation with Refined Fact Extraction and Reference Facts. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 17908–17925. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Luo, Z.; Xie, Q.; and Ananiadou, S. 2023. ChatGPT as a Factual Inconsistency Evaluator for Text Sum-marization. ArXiv preprint. arXiv:2303.15621.

Muhanad, A.; Abuelezz, I.; Khan, K.; and Ali, R. 2025. On How Cialdini's Persuasion Principles Influence In-dividuals in the Context of Social Engineering: A Qualitative Study. In Barhamgi, M.; Wang, H.; and Wang, X., eds., *Web Information Systems Engineering – WISE 2024*, Lecture Notes in Computer Science 15438, 373–388. Singapore: Springer Nature Singapore.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. *36th Conference on Neural Information Processing Systems (NeurIPS)*.

Peters, U.; and Chin-Yee, B. 2025. Generalization Bias in Large Language Model Summarization of Scientific Research. *Royal Society Open Science*, 12(4): 241776.

Pohl, R. F., ed. 2004. *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*. New York: Psychology Press.

Sarathy, V.; Burstein, M.; Friedman, S.; Bobrow, R.; and Kuter, U. 2022. A Neuro-Symbolic Cognitive System for Intuitive Argumentation. In *Proceedings of the Tenth Annual Conference on Advances in Cognitive Systems*.

Shaki, J.; Kraus, S.; and Wooldridge, M. 2023. Cognitive Effects in Large Language Models. *ECAI 2023: 26th European Conference on Artificial Intelligence*, 372: 2105–2112.

Singh, S. U.; and Namin, A. S. 2025. The In-fluence of Persuasive Techniques on Large Language Models: A Scenario-Based Study. *Computers in Human Behavior: Ar-tificial Humans*, 6: 100197.

Sonowal, H.; and Sadhu, S. 2025. Structure-Aware Chunking for Abstractive Summarization of Long Legal Documents. In *Proceedings of the 1st Workshop on NLP for Em-powering Justice (JUST-NLP 2025*, 171–178. Mumbai, India: Association for Computational Linguistics.

Stede, M.; Afantenos, S. D.; Peldszus, A.; Asher, N.; and Perret, J. 2016. Parallel Discourse Annotations on a Corpus of Short Texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, volume LREC'16, 1051–1058. Portorož, Slovenia: Euro-pean Language Resources Association (ELRA.

Suh, S.; Min, B.; Palani, S.; and Xia, H. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th An-nual ACM Symposium on User Interface Software and Tech-nology*. New York: Association for Computing Machinery.

Tam, D.; Mascarenhas, A.; Zhang, S.; Kwan, S.; Bansal, M.; and Raffel, C. 2023. Evaluating the Factual Consistency of Large Language Models Through News Sum-marization. In *Findings of the Association for Computational Linguistics: ACL 2023*, 5220–5255. Toronto, Canada: Associa-tion for Computational Linguistics.

Tang, L.; Sun, Z.; Idnay, B.; Nestor, J. G.; Soroush, A.; Elias, P. A.; Xu, Z.; Ding, Y.; Durrett, G.; Rousseau, J. F.; Weng, C.; and Peng, Y. 2023. Evaluating Large Language Models on Medical Evidence Summarization. *NPJ Digital Medicine*, 6(1): 158.

Tversky, A.; and Kahneman, D. 1974. Judgment un-der Uncertainty: Heuristics and Biases. *Science*, 185(4157): 1124–1131.

Vijay, S.; Priyanshu, A.; and KhudaBukhsh, A. R. 2025. When Neutral Summaries Are Not That Neutral: Quantifying Political Neutrality in LLM-Generated News Summaries (Student Abstract. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(28): 29514–29516.

Wang, Y.; Reddy, R. G.; Muja-hid, Z. M.; Arora, A.; Rubashevskii, A.; Geng, J.; Afzal, O.-m. M.; Pan, L.; Borenstein, N.; Pillai, A.-i.; Augenstein, I.; Gurevych, I.; and Nakov, P. 2024. Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-Checkers. *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Weick, K. E. 1995. *Sensemaking in Organizations*. Thousand Oaks, CA: Sage.