Beyond Accuracy: Forensic Evaluation of Trust and Grounding in LLM Outputs

Project Comprehension: Forensic Evaluation of LLM Outputs

A human-centered framework for analyzing plausibility, uncertainty, grounding, and actionability in high-stakes LLM decision support

Christopher Watson

Howard University, The Institute for Human-Centered Artificial Intelligence at Howard, christopher.watson@howard.edu

Janelle Yankey

Howard University, The Institute for Human-Centered Artificial Intelligence at Howard,, janelle.yankey@howard.edu

Jaye Nias, PhD.

Howard University, The Institute for Human-Centered Artificial Intelligence at Howard, jaye.nias@howard.edu

Saurav Aryal PhD.

Howard University, The Institute for Human-Centered Artificial Intelligence at Howard, saurav.aryal@howard.edu

Jeremy Blackstone, PhD.

Howard University, The Institute for Human-Centered Artificial Intelligence at Howard, jeremy.m.blackstone@howard.edu

Simone Smarr, PhD.

Howard University, The Institute for Human-Centered Artificial Intelligence at Howard, simone.smarr@howard.edu

Lucretia Williams, PhD.

Howard University, The Institute for Human-Centered Artificial Intelligence at Howard, lucretia.williams1@howard.edu

Gloria Washington, PhD.

Howard University, The Institute for Human-Centered Artificial Intelligence at Howard, gloria.washington@howard.edu

Large language models (LLMs) are increasingly used in high-stakes decision support to summarize situations, propose actions, and communicate rationale. While these systems often produce fluent and plausible responses, such outputs can obscure uncertainty, weaken grounding, and invite over-reliance by human decision-makers.

We present Project Comprehension, a forensic evaluation framework that examines LLM outputs as decision-relevant artifacts rather than isolated answers. The framework combines operationally grounded scenarios with human-centered annotation to assess plausibility, uncertainty signaling, grounding transparency, comprehension support, and actionability.

Across empirical testing, we find that surface-level response quality is only weakly predictive of grounding transparency: a non-trivial subset of responses appear clear and actionable while providing limited justification or source signaling. These patterns highlight an interpretive risk that is not captured by accuracy-focused evaluation alone.

We discuss how forensic evaluation can support trust calibration, assurance practices, and the design of language-enabled decision support systems that better align with human judgment in high-stakes contexts.

CCS CONCEPTS • Artificial Intelligence, Natural language processing, Decision support systems, Human-centered computing, Empirical studies in HCI

Additional Keywords and Phrases: LLM Forensics, Tactical decision support, Trustworthy AI, Human autonomy teaming, uncertainty communication, evaluation methodology

REFERENCES

[1] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. 2020. Language Models Are Few-Shot Learners. In Advances in Neural Information Processing Systems (NeurIPS 2020).

[2] Cambria, Erik, Leonardo Malandri, Fabio Mercorio, Nima Nobani, and Andrea Seveso. 2024. XAI Meets LLMs: A Survey of the Relation Between Explainable AI and Large Language Models. arXiv preprint arXiv:2407.15248.

[3] Clark, Herbert H., and Susan E. Brennan. 1991. Grounding in Communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley (Eds.), Perspectives on Socially Shared Cognition, 127–149. American Psychological Association, Washington, DC.

[4] Doshi-Velez, Finale, and Been Kim. 2017. Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.

[5] Gehrmann, Sebastian, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. Journal of Artificial Intelligence Research.

[6] Gigerenzer, Gerd, Wolfgang Gaissmaier, Elke Kurz-Milcke, Lisa M. Schwartz, and Steven Woloshin. 2007. Helping Doctors and Patients Make Sense of Health Statistics. Psychological Science in the Public Interest, 8(2), 53–96.

[7] Grice, H. Paul. 1975. Logic and Conversation. In Speech Acts, 41–58. Brill.

[8] Horvitz, Eric. 1999. Principles of Mixed-Initiative User Interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99), 159–166. ACM.

[9] Huang, Lei, Weiming Yu, Weijian Ma, Wenxiang Zhong, Zhiyuan Feng, Haowei Wang, Qian Chen, Wenpeng Peng, Xuanjing Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems, 43(2).

[10] Hullman, Jessica, Paul Resnick, and Eytan Adar. 2015. Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences About Reliability of Variable Ordering. PLOS ONE, 10(11), e0142444.

[11] Hutchins, Susan G., John G. Morrison, and Robert T. Kelly. 1996. Principles for Aiding Complex Military Decision Making. Technical Report. U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

[12] International Organization for Standardization. 2018. ISO 9241-11:2018 Ergonomics of Human-System Interaction – Part 11: Usability: Definitions and Concepts. ISO, Geneva, Switzerland.

[13] Ji, Ziwei, Nayeon Lee, Rita Frieske, Tianyi Yu, Dan Su, Yan Xu, Etsuko Ishii, Young Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12).

[14] Jiang, Zhengbao, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. Transactions of the Association for Computational Linguistics.

[15] Lee, John D., and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. Human Factors, 46(1), 50–80.

[16] Louvieris, Panagiotis, Andreas Gregoriades, and Warren Garn. 2010. Assessing Critical Success Factors for Military Decision Support. Expert Systems with Applications, 37(12), 8229–8241.

[17] Luger, Ewa, and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16), 5286–5297. ACM.

[18] Miller, Tim. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence, 267, 1–38.

[19] Morrison, John G., Robert T. Kelly, Robert A. Moore, and Susan G. Hutchins. 1996. Tactical Decision Making Under Stress (TADMUS) Decision Support System.

[20] National Academies of Sciences, Engineering, and Medicine. 2022. State of the Art and Research Needs. The National Academies Press, Washington, DC.

[21] Probasco, Elizabeth S., Helen Toner, Mark Burtell, and Thomas G. J. Rudner. 2025. AI for Military Decision-Making: Harnessing the Advantages and Avoiding the Risks. Technical Report. Center for Security and Emerging Technology (CSET), Georgetown University, Washington, DC.

[22] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. ACM.

[23] Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, 59–68. ACM.

[24] Steyvers, Marius, and Akash Kumar. 2024. Three Challenges for AI-Assisted Decision-Making. Perspectives on Psychological Science, 19(5), 722–734.

[25] Uziel, Stephen J. 2020. AI-Augmented Decision Support Systems: Application in Maritime Decision Making Under Conditions of METOC Uncertainty. Ph.D. Dissertation. Naval Postgraduate School, Monterey, CA.