LLMs Need to Go Beyond Computational Confidence Metrics to Establish Trust

Anil B Murthy[1], **Lindsay Sanneman**[1]

[1]School of Computing & Augmented Intelligence, Arizona State University
abmurthy@asu.edu, lindsay.sanneman@asu.edu

Abstract

While Large Language Models (LLMs) have demonstrated impressive capabilities, their widespread deployment is hindered by the lack of trustworthiness of their responses. Although existing trust scores and confidence metrics attempt to quantify the uncertainty, ensure safety, and reliability of LLM responses, they address only a single dimension of trust and fail to ensure trust holistically, in a user-centric manner. This lack of metric reliability and LLM trustworthiness poses significant risks in critical human-AI interaction applications. We posit that current confidence metrics and trust scores are insufficient to accurately measure trustworthiness and to ultimately inform how to establish calibrated user trust in these systems. We further argue that we need to move beyond computational assessments to enhance the measurement of trustworthiness of generative AI systems. We outline frameworks and approaches that can be incorporated into holistic trustworthy AI assessment and development in future research.

Introduction

Large Language Models(LLMs) have been found to be useful for a plethora of applications, from question answering and document summarization to code generation. Despite their utility and web-scale training, there are significant concerns about the reliability and trustworthiness of their responses. Due to the autoregressive nature of these black-box models, their responses are based on a probabilistic process of token-by-token generation derived from the vast amount of internet data that they are trained on. While existing works have focused on computing trust scores, confidence metrics, and quantifying the uncertainty of LLM responses (Chen and Mueller 2024; Kadavath et al. 2022; Tian et al. 2023; Lin, Trivedi, and Sun 2024), less emphasis has been placed on user-centric evaluations of LLM trustworthiness. Existing trust scores and confidence metrics account for the computational models of trust, but most do not address other dimensions and characteristics of trust that are necessary to ensure appropriate calibration of user trust in LLMs.

In this paper, we discuss the dimensions and characteristics of trust from the perspective of human-AI interaction, examine the existing literature on proposed confidence metrics, uncertainty measures, and trust scores, and identify significant gaps in evaluating and establishing the trustworthiness of generative AI systems in a human-centered framework. We further propose approaches, methods, and frameworks that can be incorporated to address the gaps with respect to additional human-centered dimensions of trust, with the ultimate aim of enhancing approaches to develop trustworthy and transparent generative AI systems.

We begin by presenting a detailed description of the definitions, bases, and characteristics of trust. Then, we dive deep into the technical methods and derivations of existing state-of-the-art confidence metrics and trust scores, and highlight their shortcomings with respect to the dimensions and characteristics of trust. Finally, we propose approaches and frameworks that can be incorporated to address the missing dimensions of trust and conclude with a summary of the position and need for future research in this direction.

Trust: Definitions, Dimensions and Characteristics

The concept of trust has been widely studied with various established perspectives from multiple fields. One of the most widely used and accepted perspectives in Human-AI interaction is the definition of trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee and See 2004). Lee and See further define 'purpose', 'process', and 'performance' as three bases of trust that define the types of information a user needs in order to maintain an appropriate level of trust in an automated system. **As LLMs are increasingly being adopted by users for generic search and other daily assistive tasks, we argue that the assessment of the trustworthiness of these systems must go beyond evaluations of system properties or their output indicators, and include human-centered evaluations along these dimensions as well**.

"Purpose" is defined as the degree to which an automated system (agent) is being used within the realm of the designer's intent (Lee and See 2004). This means that a user would need to know what capabilities the agent has been designed to perform. For example, a user interacting with an LLM needs to know what kind of questions can and cannot be answered by the system. "Process" refers to the degree to which the agent's algorithm is appropriate for the

situation and achieves the user's goals. Lee and See emphasize that with "process" information, the user's trust is in the agent itself rather than in specific actions taken by the agent. In essence, the process basis of trust refers to the algorithms and operations that govern the behavior of the agent, and not the behavior itself. This implies that a user would need to know approximately how an agent performs its tasks. For example, a user interacting with an LLM-based application needs to know whether it is retrieving information from the Internet or generating answers from memory. "Performance" refers to the agent's demonstrated operations and includes characteristics such as reliability, predictability, and ability (Lee and See 2004). Thus, performance information relates to the competency or expertise of the agent as demonstrated by its ability to achieve the user's goals. This basis of trust is task and situation-dependent. With performance information, a user would obtain knowledge of the agent's reliability, predictability, or ability within the context or domain of interest. For example, an LLM that provides a confidence score with each answer conveys performance information to the user.

Trust calibration refers to the correspondence between a user's trust in the agent and the agent's capabilities (Lee and Moray 1994; Muir 1987). It is necessary to establish an appropriately calibrated level of trust in the agent to avoid misuse or disuse among users (Lee and See 2004). Overtrust is poor calibration in which trust exceeds the agent's capabilities and can lead to overuse or automation bias (defined as over-reliance on automated systems even when their output contradicts accurate human judgement) (Mosier et al. 1996, 1997; Parasuraman and Riley 1997). On the other hand, undertrust can lead to disuse and algorithm aversion (Dietvorst, Simmons, and Massey 2015; Dawes 2008). Trust calibration can be achieved by providing (additional) information, explanations, or making agent behavior more interpretable, consistent, or transparent (Sanneman and Shah 2020; Wang, Pynadath, and Hill 2016; Schweitzer, Hershey, and Bradlow 2006; Sanneman and Shah 2022).

Prior research also emphasizes that trust has two key properties, namely vulnerability and anticipation (Jacovi et al. 2021). Vulnerability, as also mentioned above in Lee and See's definition of trust, arises from users' dependence on the agent's decisions and capabilities, especially when outcomes affect the users' goals or well-being (Jacovi et al. 2021; Zahedi 2023). In other words, users' acceptance of vulnerability to an agent's actions characterizes their trust in the agent, and similarly, the contrapositive statement holds, i.e., refusal characterizes distrust in the agent. For example, a patient providing personal health records to a medical AI application and prompting it for a diagnosis (informally or implicitly) accepts vulnerability to the application's diagnosis result or decision, which derives from her trust in the application. Some definitions portray trust as an outcome of behavior or as a state of vulnerability (Meyer 2001).

Whereas anticipation refers to the user's ability to anticipate the impact of an agent's decisions, and hence is an expectation about future behavior (Jacovi et al. 2021). Pragmatically, 'anticipating' refers to a belief that the trustee (agent) will act in the trustor's (user's) best interests. Also, trust can be thought of as an attempt to anticipate the impact of the agent's behavior under risk, uncertainty, or vulnerability (Hoffman 2017). Thus, the user's anticipation and trust can change as they gain more information about the agent's competence, reliability, confidence, and other attributes (Zahedi 2023). Thus, the three information types - purpose, process, and performance- can influence the two key properties of trust - vulnerability and anticipation (Jacovi et al. 2021). Further, these two critical properties of trust - vulnerability of a user in the face of decision-making AI systems and anticipation (human's psychological expectancy) of AI's actions and potential impact- embody the essence of trust and underscore its significant role in the context of human-AI interaction applications (Zahedi 2023).

LLMs' Confidence Metrics and Trust Scores

Recent literature has centered on proposing computational trust metrics to estimate the trustworthiness of LLM responses, relying purely on computational evaluation of system properties or their output indicators. These methods span from directly prompting models for confidence estimates to using similarity measures, sampling-based approaches, and leveraging LLMs' internal logit/ token-level generation probabilities (Wang et al. 2023; Zhou et al. 2025). Here, we discuss a few commonly cited metrics in depth.

Chen and Mueller (Chen and Mueller 2024) introduced BSDetector, which produces a confidence score for every LLM response. The confidence score is estimated from two factors: Observed Consistency (O) and Self-reflection Certainty (S), which are proposed as extrinsic and intrinsic measures, respectively. Observed consistency is a sampling-based measure, where 'k' responses are sampled from the LLM by varying its temperature parameters and then computing semantic similarities between the samples and the original response. Self-reflection Certainty (S) asks the model to reflect on its response (given the prompt + question + response) and provide a rating on a multiple-choice scale with 'correct', 'wrong', and 'not sure' options - which correspond to numerical scores of 1, 0, and 0.5, respectively. The overall confidence score (C) is estimated by aggregating these two factors with a tradeoff parameter $\beta \in (0, 1)$, which has lower values for LLMs that are found to be more trustworthy in their self-reflection capabilities, as shown in Eq. 1.

$$C = \beta O + (1 - \beta)S \tag{1}$$

Utilizing the consistency assumption of sampled LLM responses, SelfCheckGPT (Manakul, Liusie, and Gales 2023) detects hallucinated facts. This works on the idea that if the question is within LLMs' knowledge (pretraining data distribution), then sampled answers are likely to be similar and contain consistent facts. However, for hallucinated facts, the stochastically sampled responses are likely to diverge and contradict one another. As checking the consistency of sampled responses is a post-hoc processing method and does not require token-level probability information, SelfCheckGPT is a black-box system, purely assessing the probability of hallucinations through consistency methods. The consistency checking is performed using multiple techniques

such as BERTScore, Natural Language Inference, Multiple-choice answers, n-gram models, and direct prompting. It is important to note that, when multiple answers or reasoning paths are available for a given problem, self-consistency and self-reflection values are not reliable (Wang et al. 2023; Huang et al. 2023; Renze and Guven 2024). For example, a vague question may automatically have multiple satisficing but inconsistent answers. An out-of-distribution complex question may elicit a smaller number of (unique) responses even when temperature is varied, compared to other questions.

Semantic entropy (Kuhn, Gal, and Farquhar 2023) provides an estimate of confidence levels of LLM responses by incorporating linguistic invariances created by shared meanings. The linguistic invariances are generated by sampling from the predictive distribution of an LLM given a context, and then clustering them through bidirectional entailment. Then the resulting entropy is estimated by summing the probabilities of sequences that share a meaning. The approach also utilizes LLMs' internal token-level probabilities for computing semantic equivalence.

There are other metrics proposed to estimate the uncertainty and confidence of LLM responses from similarity measures, such as the sum of eigenvalues of the graph Laplacian, which takes into account the number of semantic meanings rather than simple semantic equivalence to quantify uncertainty (Lin, Trivedi, and Sun 2024). This metric uses the distribution of eigenvalues in spectral clustering of LLM responses to determine the number of clusters, which roughly corresponds to the number of semantic meanings. Lin, Trivedi, and Sun further propose the degree matrix method, where they use a degree matrix obtained in the intermediate steps of the eigenvalue graph Laplacian method to compute a confidence as well as an uncertainty score, by noting that nodes with a higher degree of connections correspond to more confident regions of the LLM.

LLMs can also be directly prompted to provide confidence scores, similar to the self-reflection certainty factor mentioned above. LLMs are asked to first propose an answer and then prompted to evaluate the probability "P(True)" (Kadavath et al. 2022). Kadavath et. al. also find that models can perform well at predicting the probability P("I Know") given a question and prompt without reference to any proposed answer, and that models partially generalize across tasks but struggle to calibrate the probability on new tasks. Similarly, Lin, Hilton, and Evans (2022) ask LLMs for "verbalized probability", expressing uncertainty in words, expecting it to be low when it is likely to get the answer wrong. They finetune GPT-3 on a calibrated Math dataset to produce verbalized probabilities, finding that it has some ability to generalize calibration under distribution shift. Finally, Tian et. al. (Tian et al. 2023) finetune an LLM using Reinforcement learning with Human Feedback (RLHF) and elicit calibrated confidences by prompting the model to verbalize its confidence in token space. They find that the verbalized confidences are better calibrated than LLMs' internal conditional probabilities across several closed-source models.

Each of these confidence measures is obtained either by direct prompting or by computing score estimates based on additional metrics such as sampling or token probabilities, but they are still not robust and accurate enough to be reliable. As certain questions, such as "What are the different types of ML algorithms?" are intrinsically highly disjunctive in possible correct answers or otherwise have multiple plausible answers or reasoning paths, confidence score estimates based on sampling, consistency, self-reflection, and token probabilities can vary widely, leading to inaccuracies and unreliable scores. In spite of these shortcomings, these metrics do provide information related to the model's performance, confidence, and reliability, which can be helpful to the user. Although re-prompting models to correct their answers after self-reflection or additional feedback has shown promise, these strategies still do not provide any irrefutably correct or reliably accurate responses at all times (Huang et al. 2023; Renze and Guven 2024). For example, it has been shown that self-verification abilities of LLMs fare poorly for problems such as logical reasoning and planning (Hong et al. 2023; Stechly, Valmeekam, and Kambhampati 2024). Therefore, evaluating the trustworthiness of a particular LLM must go beyond self-evaluated confidence scores or consistency-related metrics, as discussed next.

Relation with Trust Dimensions and Characteristics

All the above works propose methods that elicit confidence scores either directly from the model or compute an estimated score empirically, utilizing metrics such as similarity measures, self-consistency, self-reflection, or token-level probabilities. These methods address the performance basis of trust as they provide information relating to the model's (agent's) competence, confidence, performance, or reliability of answers or decisions. This performance information, when reliable and accurate, can perhaps help users gauge the confidence and reliability of the model's ability to answer questions and even anticipate the nature of its responses to certain questions. Where existing metrics suffer in accuracy or reliability, alternate performance-related metrics must be employed. In addition, the information relating to the other two bases of trust, namely purpose and process, is missing within the scope of existing trustworthiness metrics and evaluations, which can affect the characteristics of user trust.

Although confidence metrics and trust scores are representative of performance information, their reliability in reasoning problems may hinder the trust of the user. As LLMs are known to struggle at simple reasoning tasks such as counting, poor confidence scores for obviously correct answers can greatly affect trust (Fu et al. 2024). It has been demonstrated and validated through user studies that for reasoning problems such as planning, correctness is the primary driver of trust and performance (Chen et al. 2025). Then, to ensure correctness and, therefore, increase trust, formal verification through automated domain-specific verifiers, tool-calling, and mitigating hallucinations through retriever-aware training are useful ways forward for reasoning problems (Kambhampati et al. 2024; Patil et al. 2024; Gao et al. 2025). And confidence metrics can be used in this pursuit to help decide on tool-calling, verification, and model selection between LLMs and reasoning models.

Missing Dimensions of Trust: What Can Be Done to Improve Trustworthiness?

One of the crucial aspects of an AI system's trustworthiness is the transparency of its capabilities. Although LLMs, especially closed-source models, have been largely uninterpretable in terms of their capabilities, there are still numerous ways to communicate information such as data, usage policies, risks, model access, and other dimensions of the Foundation Model Transparency Index (Maslej et al. 2025), to enhance the interpretability and transparency of the system as shown in Figure 1.

One such way is abstention (defined as the refusal to answer a query), which has been demonstrated to enhance model safety and reliability. However, the progress is still in the nascent stages, requiring abstention mechanisms to be more adaptive and context-aware so that AI systems are more robust, reliable, and trustworthy (Wen et al. 2025). We note that a model demonstrating abstinence also conveys "purpose" related information that provides a user with useful information on the limits of the model's capabilities. Another similar type of information that is useful for the "purpose" dimension in LLM-based applications is usage policies, which convey the terms of recommended use and non-use of the application. Similarly, the communication of "Operational Design Domain" of a generative AI application in deployment, which defines the specific operating conditions under which the application is designed to function, as defined for autonomous vehicles (NHTSA 2017), is useful for the user to learn "purpose"-related information.

Concerning the "Process" basis of trust, information on the algorithms that govern the model's behavior that can be considered to be provided to users (through model information cards, training data summaries, or model documentation forms (European Commission 2025)) includes answers to the following questions:

- Is the model or application connected to any external tools, verifiers, validators, or other (overseer) models?
- Is the model retrieving information from the internet in real-time or generating responses from memory?
- Are responses being retrieved from any public, private, or enterprise databases? Are citations within responses publicly accessible and open-source?
- What is the nature, constitution, and curation framework of the model's training data?
- How is the model trained? For example, is it a reasoning model or an autoregressive model? Does the model employ routing (Yue et al. 2024) or inference-time scaling techniques (Wu et al. 2024) for any questions?

Alongside computing the confidence and uncertainty metrics, another aspect that can be useful is to provide confidence and capability warnings alongside responses, such as a note that says "Response is not verified", if no verification is involved or "This response may be considered toxic", if toxicity evaluations are performed. This information relates to the "process" basis of trust. As LLMs are being continuously benchmarked on various capabilities, a useful evaluation-related information for better transparency can be stating that the particular model has an "X% accuracy on Mathematical Questions" or "Y% accuracy on Medical Questions" or similar (Reuel-Lamparth et al. 2024). This kind of benchmark evaluation information constitutes the "purpose" basis of trust and can be useful for trust calibration. Users' trust in the LLMs they are interacting with needs to be calibrated and aligned with the level of the systems' capabilities. Both of such types of information can be similarly provided to the users as community notes on social media platforms, which have been introduced to enhance transparency, safety, and reduce harm (X Corp. 2025).

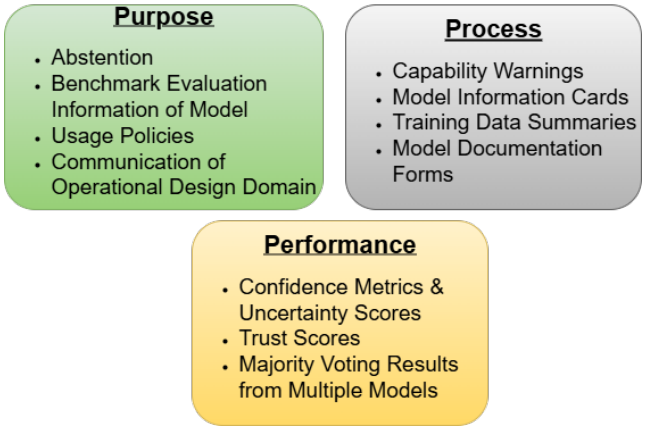

Figure 1: Information Framework for Trust Dimensions

While the above proposed approaches provide a solid foundation for future research, a promising experiment would be to evaluate the trustworthiness of generative AI systems enhanced with purpose, process, and performance information holistically (i.e. using the computational metrics mentioned, provided that they are made to be reliable through verification or knowledge retrieval), along with human-centered evaluations on the trustworthiness of the systems' responses and effectiveness of trust calibration. Evaluating generative AI systems incorporated with trust considerations could pave the way for building better, more transparent, and trustworthy AI applications for the future.

Conclusion

In this paper, we have argued that current metrics to estimate confidence and trustworthiness of LLM responses address only the performance basis of trust, and that we need to go beyond the computational aspects to establish trust. In support of this position, we first detailed the definitions, characteristics, and bases of trust from a human-AI interaction perspective. Then, we reviewed current confidence, uncertainty, and trust score metrics from the literature, and discussed results that demonstrate the unreliability and lack of robustness of the stated metrics, and how they do not address the gaps within the "purpose" and "process" dimensions of trust. Then, we shared the possible approaches, methods, and frameworks that can be incorporated to address the gaps within the missing dimensions of trust, calling for further research and user studies to enhance the trustworthiness, transparency, and safety of LLMs and their applications.

References

Chen, J.; and Mueller, J. 2024. Quantifying Uncertainty in Answers from any Language Model and Enhancing their Trustworthiness. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5186–5200. Bangkok, Thailand: Association for Computational Linguistics.

Chen, S.; Yang, Y.; Boggess, K.; Heo, S.; Feng, L.; and Topcu, U. 2025. Evaluating human trust in llm-based planners: A preliminary study. *arXiv preprint arXiv:2502.20284*.

Dawes, R. M. 2008. The robust beauty of improper linear models in decision making. In *Rationality and social responsibility*, 321–344. Psychology Press.

Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, 144(1): 114.

European Commission. 2025. Code of Practice on general-purpose AI. European Union Policy Document. Accessed: September 2025.

Fu, T.; Ferrando, R.; Conde, J.; Arriaga, C.; and Reviriego, P. 2024. Why Do Large Language Models (LLMs) Struggle to Count Letters? *arXiv preprint arXiv:2412.18626*.

Gao, S.; Dwivedi-Yu, J.; Yu, P.; Tan, X. E.; Pasunuru, R.; Golovneva, O.; Sinha, K.; Celikyilmaz, A.; Bosselut, A.; and Wang, T. 2025. Efficient Tool Use with Chain-of-Abstraction Reasoning. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 2727–2743. Abu Dhabi, UAE: Association for Computational Linguistics.

Hoffman, R. R. 2017. A taxonomy of emergent trusting in the human–machine relationship. *Cognitive Systems Engineering*, 137–164.

Hong, R.; Zhang, H.; Pang, X.; Yu, D.; and Zhang, C. 2023. A Closer Look at the Self-Verification Abilities of Large Language Models in Logical Reasoning.

Huang, J.; Chen, X.; Mishra, S.; Zheng, H. S.; Yu, A. W.; Song, X.; and Zhou, D. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Jacovi, A.; Marasović, A.; Miller, T.; and Goldberg, Y. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 624–635.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kambhampati, S.; Valmeekam, K.; Guan, L.; Verma, M.; Stechly, K.; Bhambri, S.; Saldyt, L.; and Murthy, A. 2024. Position: LLMs can't plan, but can help planning in LLM-modulo frameworks. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Lee, J. D.; and Moray, N. 1994. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40(1): 153–184.

Lee, J. D.; and See, K. A. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1): 50–80. PMID: 15151155.

Lin, S.; Hilton, J.; and Evans, O. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Lin, Z.; Trivedi, S.; and Sun, J. 2024. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *Transactions on Machine Learning Research*.

Manakul, P.; Liusie, A.; and Gales, M. J. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Maslej, N.; Fattorini, L.; Perrault, R.; Gil, Y.; Parli, V.; Kariuki, N.; Capstick, E.; Reuel, A.; Brynjolfsson, E.; Etchemendy, J.; et al. 2025. Artificial intelligence index report 2025. *arXiv preprint arXiv:2504.07139*.

Meyer, J. 2001. Effects of Warning Validity and Proximity on Responses to Warnings. *Human Factors*, 43(4): 563–572. PMID: 12002005.

Mosier, K. L.; Skitka, L. J.; Burdick, M. D.; and Heers, S. T. 1996. Automation Bias, Accountability, and Verification Behaviors. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 40(4): 204–208.

Mosier, K. L.; Skitka, L. J.; Heers, S.; and Burdick, M. 1997. Automation bias: decision making and performance in high-tech cockpits. *Int. J. Aviat. Psychol.*, 8(1): 47–63.

Muir, B. M. 1987. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27(5): 527–539.

NHTSA. 2017. Automated driving systems 2.0: A vision for safety. *Washington, DC: US Department of Transportation, DOT HS*, 812: 442.

Parasuraman, R.; and Riley, V. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, 39(2): 230–253.

Patil, S. G.; Zhang, T.; Wang, X.; and Gonzalez, J. E. 2024. Gorilla: Large language model connected with massive apis. *Advances in Neural Information Processing Systems*, 37: 126544–126565.

Renze, M.; and Guven, E. 2024. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*.

Reuel-Lamparth, A.; Hardy, A.; Smith, C.; Lamparth, M.; Hardy, M.; and Kochenderfer, M. J. 2024. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems*, 37: 21763–21813.

Sanneman, L.; and Shah, J. A. 2020. Trust considerations for explainable robots: A human factors perspective. *arXiv preprint arXiv:2005.05940*.

Sanneman, L.; and Shah, J. A. 2022. The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems. *International Journal of Human–Computer Interaction*, 38(18-20): 1772–1788.

Schweitzer, M. E.; Hershey, J. C.; and Bradlow, E. T. 2006. Promises and lies: Restoring violated trust. *Organizational behavior and human decision processes*, 101(1): 1–19.

Stechly, K.; Valmeekam, K.; and Kambhampati, S. 2024. On the self-verification limitations of large language models on reasoning and planning tasks. *arXiv preprint arXiv:2402.08115*.

Tian, K.; Mitchell, E.; Zhou, A.; Sharma, A.; Rafailov, R.; Yao, H.; Finn, C.; and Manning, C. D. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Wang, N.; Pynadath, D. V.; and Hill, S. G. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 109–116. IEEE.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Wen, B.; Yao, J.; Feng, S.; Xu, C.; Tsvetkov, Y.; Howe, B.; and Wang, L. L. 2025. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13: 529–556.

Wu, Y.; Sun, Z.; Li, S.; Welleck, S.; and Yang, Y. 2024. Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for LLM Problem-Solving. In *International Conference on Learning Representations*.

X Corp. 2025. Introduction to Community Notes. https://communitynotes.x.com/guide/en/about/introduction. Accessed on August 11, 2025.

Yue, M.; Zhao, J.; Zhang, M.; Du, L.; and Yao, Z. 2024. Large Language Model Cascades with Mixture of Thought Representations for Cost-Efficient Reasoning. In *The Twelfth International Conference on Learning Representations*.

Zahedi, Z. 2023. *Computational Accounts of Trust in Human AI Interaction*. Ph.D. thesis, Arizona State University.

Zhou, Z.; Yuhao, T.; Li, Z.; Yao, Y.; Guo, L.-Z.; Ma, X.; and Li, Y.-F. 2025. Bridging internal probability and self-consistency for effective and efficient llm reasoning. *arXiv preprint arXiv:2502.00511*.