Research

Measuring the Impact of Hallucinations on Human Reliance in LLM Applications

Mohammad Hassan[1]

[1] Dhaka, Bangladesh

Abstract: Modern large language models generate outputs that often exhibit unexpected or fabricated details, commonly referred to as hallucinations, influencing how humans interpret and rely upon these systems. Behavioral experiments show that users sometimes defer to system outputs, assuming correctness in contexts where thorough verification may not be feasible. Recent studies highlight that such misplaced reliance can manifest in high-stakes domains, including medical triage, legal documentation, and policy recommendations, where the costs of erroneous information are severe. Quantitative assessments typically gauge hallucinations in terms of factual inconsistencies, yet the downstream human impact remains less systematically investigated. This paper develops an evaluation pipeline that measures the extent to which hallucinations shape user decisions and reliability judgments. By integrating controlled prompts with varied levels of fidelity, the approach isolates the effects of erroneous content from user-specific biases. Empirical results present evidence that even low-frequency hallucinations can erode trust and lead to suboptimal task performance in collaborative human–machine settings. This finding shows the importance of accurate metrics that capture not only the presence of factual deviations but also their effects on human behavior. A deeper understanding of these behavioral dimensions can inform the design of guidelines and protocols aimed at maintaining user engagement without inflating unwarranted trust in generative outputs.

1. Introduction

Large language models (LLMs) have become central to a wide range of applications, from summarizing scholarly articles and generating product descriptions to assisting in technical problem-solving and creative writing. Ongoing advancements in training architecture and data curation have driven consistent improvements in semantic coherence, contextual accuracy, and fluency of generated text. Many systems now leverage multi-billion-parameter configurations to capture nuanced linguistic patterns, enabling interactions that appear to exhibit human-like depth and responsiveness. Empirical tests confirm the capacity of these models to handle diverse tasks, yet the phenomenon of hallucination remains a persistent challenge. Hallucinations materialize when a model produces output that deviates from verifiable reality, leading to factual inaccuracies or invented references. Such deviations present a risk for users who may lack the means or time to perform detailed verifications.

User reliance on LLM-generated content often hinges on perceived credibility, seamlessness of interaction, and a user's subjective trust in computational competence [1,2]. Evidence suggests that humans sometimes place undue weight on content delivered through convincingly articulated channels, attributing authority and expertise to systems that seem capable of sophisticated reasoning. This dynamic becomes acute when the user's domain expertise is limited. The question of how hallucinations reshape these trust dynamics remains underexplored in both methodological rigor and empirical scale. System developers focus on optimizing language model parameters to minimize hallucinations,

. . *Helex-science* **2025**, *10*, 10–20.

Factor	Description	Impact on Output	Example
Training Data Bias	Imbalanced or low-quality sources	Reinforces inaccuracies	Overrepresentation of certain perspectives
Overgeneralization	Excessive reliance on patterns	Generates plausible but incorrect facts	Assigning wrong authorship to papers
Incomplete Context Processing	Failure to capture nuanced dependencies	Omits critical information or distorts meaning	Misinterpretation of ambiguous queries
Decoder Sampling Strategies	Choice of token selection mechanisms	Influences fluency vs. factual accuracy trade-off	Higher hallucination risk with nucleus sampling

Table 1. Key factors that contribute to hallucinations in large language models (LLMs).

but understanding the exact impact on human reliance requires nuanced behavioral studies [3].

Human reliance operates at the intersection of cognitive psychology, user interface design, and risk assessment. Certain tasks demand a high level of scrutiny due to potentially severe consequences of errors, as in automated legal document generation or healthcare decision support. Even in domains where the stakes are relatively modest, misrepresentations can accumulate over time to produce far-reaching repercussions. For instance, repeated reliance on errant model outputs in an educational setting could foster misconceptions that persist, undermining knowledge acquisition. An emphasis on cursory validation or superficial checks may fail to detect deep-seated fabrications embedded within otherwise coherent prose.

Quantitative analysis of hallucinations frequently focuses on matching generated text to ground-truth references. Automated metrics attempt to measure divergences from known facts, yet the user-side effects are not strictly captured by these techniques. Some responses containing partial inaccuracies might still carry enough useful information to guide a user's reasoning, whereas minor hallucinatory details in other contexts might lead to serious errors, driven by overreliance on perceived authority of the system. Beyond the brute-force quantification of factual mismatches, investigations must incorporate a behavioral dimension that tracks when and how humans accept or contest these outputs.

Factor	Influence on Reliance	Implications	Example Scenario
Perceived Credibility	Confidence in system accuracy	Users accept outputs without verification	Trusting a model's historical analysis
Domain Expertise	User knowledge in the subject	Higher expertise leads to critical evaluation	Experts detecting inconsistencies in legal texts
Task Urgency	Time constraints on validation	Users rely more under pressure	Quick decision-making in customer support
Transparency Mechanisms	Explanation of model reasoning	Can reduce blind reliance but not eliminate errors	Justification-based interfaces in medical AI

Table 2. Factors influencing user reliance on LLM-generated content.

Observational and survey-based studies propose that model explanations or system transparency can influence user trust, but such interventions do not eliminate the emergence of hallucinated content. Task context also plays a pivotal role: research in human–machine teaming indicates that reliance varies with time pressures, interface complexity, and the difficulty of validating specific pieces of information. These variables suggest a multifaceted approach for capturing the interplay between hallucination frequency, user trust calibration, and ultimate task success rates.

Critically assessing hallucinations and their influence on human reliance demands a research framework that integrates objective measurement of textual fidelity with subjective indices of trust and acceptance. The following sections offer a methodological blueprint, experimental protocol, and both quantitative and qualitative analyses that illuminate how hallucinations affect user reliance. By highlighting structured experiments involving controlled prompts and rigorous data collection procedures, the study endeavors to delineate the connections between erroneous content and human decision-making processes. Subsequently, the discussion delves into interpretative findings of user trust behaviors, pinpointing the extent to which hallucinations can sway individuals toward suboptimal reliance patterns [4,5].

Strategy	Implementation	Benefits	Limitations
Retrieval-Augmented Generation	Incorporates external sources at inference time	Improves factual grounding	Computationally expensive
Post-Generation Verification	Uses fact-checking modules on generated text	Reduces misinformation risk	Slows response time
Training Data Refinement	Enhancing dataset quality [6]	Lowers initial hallucination frequency	Requires ongoing curation
Confidence Calibration	Adjusts probability outputs to reflect uncertainty	Helps users gauge reliability	May affect fluency in outputs

Table 3. Strategies for mitigating hallucinations in large language models.

2. Methodological Framework for Evaluating Hallucinations

Methodological approaches to measuring hallucinations in LLM outputs frequently adopt automated or semi-automated systems that compare generated text to curated knowledge bases. Such methods create a benchmark of factual correctness against which a system's performance is scored. Existing approaches, while valuable, generally focus on the text itself, overlooking how users respond to hallucinatory statements in realistic scenarios. An integrated methodology must incorporate behavioral variables that capture user acceptance and decision-making under varying conditions of uncertainty.

Data collection in this study was guided by a three-tiered framework. The first tier addressed objective identification of hallucinations by aligning model outputs with authoritative references in a controlled domain. The second tier assessed user recognition or detection of errors through structured feedback mechanisms. The third tier measured behavioral changes or reliance patterns, allowing the research team to evaluate the direct effects of hallucinated statements on real-time decision-making tasks.

Initial corpus preparation required careful curation of reference materials. Domain experts in law, healthcare, and technology were recruited to construct verified factual repositories. These repositories included short passages, case examples with validated data, and scenario-based fact sheets to represent complex real-world tasks. During the corpus development phase, specialized attention was given to ensuring that reference materials covered a broad range of difficulty levels. Some topics were straightforward to verify, such as numerical facts or definitions, whereas others involved intricate reasoning or interdisciplinary knowledge that posed higher cognitive demands on users.

Controlled prompt generation formed a critical component of the methodology. Prompts were designed to elicit content from the LLM that would either be entirely factually aligned or contain targeted hallucinations. Researchers systematically manipulated prompt structures and inserted ambiguous elements to induce a spectrum of potential hallucinatory outputs. By standardizing prompt templates, the experimental design facilitated precise comparisons across different user groups and tasks. In addition, each prompt was annotated with metadata detailing domain, difficulty level, and potential areas for hallucination.

User participation was structured to reflect real-world decision contexts. Volunteers were asked to engage with the system on tasks aligned with their domain familiarity. A medical professional, for instance, might receive a scenario related to treatment guidelines, whereas a legal clerk might assess a scenario involving contractual clauses. The user interface integrated text generation from the model, followed by a question asking for a decision based on the provided content. Various trials introduced subtle illusions or overt factual fabrications, all embedded in a coherent narrative. Participants then indicated their confidence in the system output and recorded any doubts about its validity.

Behavioral data was recorded through multiple channels. First, user interactions were logged, capturing time spent reading outputs, frequency of re-checking facts, and whether participants requested clarifications. Second, immediate post-task surveys queried confidence levels and trust perceptions, using five-point scales and open-ended justifications. Third, an optional delayed feedback survey was administered to examine whether participants detected errors upon reflection. This delayed component sought to uncover shifts in reliance that might surface after users had an opportunity to revisit or verify the content independently.

Analytical rigor in evaluating the impact of hallucinations on reliance called for quantitative models of user responses. Hierarchical regression was used to estimate the relationship between hallucination presence, recognition, and reliance, controlling for user expertise, task complexity, and domain familiarity. The model distinguished between partial and full reliance, enabling a more granular understanding of how localized factual deviations could alter broader trust dynamics. Researchers interpreted effect sizes with caution, focusing on operationally relevant thresholds of user trust rather than purely statistical significance.

Selection biases were addressed by diversifying the user sample. Participants were drawn from different professional backgrounds, age brackets, and educational levels. Randomization of prompt assignments further minimized systematic differences across groups. Additionally, the study employed an independent coding team to label textual segments for potential hallucinations, thus reducing biases introduced by the primary investigators. Validation checks ensured consistency among coders, and inter-rater reliability was measured via Cohen's kappa to confirm that identified hallucinations were marked with minimal subjectivity.

Precision in measuring factual correctness involved reference matching at both the sentence level and entity level. The system outputs underwent Named Entity Recognition (NER) to identify critical items such as dates, names, and technical terms. These items were then compared against the reference set to detect discrepancies. Subject-matter experts reviewed flagged mismatches to classify them as minor or major factual deviations. Major deviations included misrepresentations of key concepts that could materially affect user decisions. Minor deviations, though potentially harmless in certain contexts, were still monitored for their cumulative effect on trust [7,8].

Integrating objective textual analysis with user feedback culminated in a dataset linking hallucination presence to subsequent reliance patterns. This dataset enabled a deeper understanding of how individuals weigh system credibility when confronted with partial inaccuracies [9]. The layered approach—tracking hallucinations from generation to user response—aimed to provide a holistic view of the phenomenon.

3. Experimental Setup and Data Collection

Participant recruitment followed an Institutional Review Board (IRB)-approved protocol to ensure ethical considerations and data privacy. A total of 300 participants were initially enrolled, encompassing various domains such as law, healthcare, software engineering, finance, and general administrative tasks. Each participant received an orientation packet that outlined the study's goals without revealing the specific focus on hallucinations. Informed consent was obtained, and participants were assured that their responses would remain anonymous.

Task sequences were randomized to reduce order effects, ensuring that no single participant or group repeatedly encountered analogous prompts. The randomization process balanced domain coverage, difficulty level, and presence or absence of induced hallucinations. Prompt sets were assembled into blocks, each containing a mixture of factual correctness levels. This arrangement prevented participants from deducing patterns that might influence their natural response behavior.

Data capture mechanisms included an online platform that integrated dynamic text generation, interactive prompts, and survey components. Each participant interacted with the LLM through a custom interface that displayed the system's responses in real time. The interface requested the participant's decision immediately after receiving the generated text, followed by confidence assessments ranging from zero (complete distrust) to four (high trust). Free-text boxes allowed participants to note any discrepancies they perceived or articulate uncertainties. Responses were automatically time-stamped for subsequent analysis.

Instrumentation extended beyond behavioral logging by incorporating eye-tracking for a subset of participants. This technology tracked the sequence and duration of visual focus on different portions of the generated text. Eye-tracking data helped reveal when individuals devoted extra attention to suspect segments, suggesting they might be identifying potential hallucinations. Comparative analyses between eye-tracking participants and the larger sample offered deeper insights but also required additional exclusion criteria for data quality. The final analysis only integrated eye-tracking data for participants who maintained valid calibration metrics throughout the session.

Qualitative interviews were conducted with a smaller subgroup to delve deeper into decision-making rationales. This subgroup included 30 participants selected to represent high, medium, and low reliance on the system. Each interview lasted approximately 30 minutes, focusing on how participants interpreted the model outputs, the reasons they accepted or questioned certain statements, and how these decisions aligned with their prior knowledge. Interview transcripts were thematically coded, revealing recurrent patterns of trust formation, skepticism triggers, and heuristic cues used by participants when appraising text quality.

Parallel to user data collection, the system's output was subjected to automated scrutiny via a pipeline that flagged potential hallucinations. A database of known facts and accepted domain practices served as a comparative ground truth. Each output was decomposed into sentences, clauses, and entities. Mismatched entities, contradictory statements, or fabricated references were flagged. A submodule examined coherence, looking for text segments that introduced inconsistent narrative elements. These flagged segments were then cross-checked by human raters with domain expertise to confirm hallucination presence and severity. Although automated detection was not foolproof, it provided a systematic pre-screening that accelerated human verification [10,11].

Data underwent preliminary processing to remove incomplete sessions, define participant-level metrics, and standardize variable naming conventions. Instances where participants did not submit confidence ratings were excluded from the final analysis. Additional filters removed data points where the system output was truncated or the interface encountered a technical glitch. Approximately 10% of the dataset was lost due to these inconsistencies [12], leaving 270 participants and around 5,000 individual decision instances across domains.

Demographic distributions were examined for potential confounders. The participant pool had an approximate gender balance of 53% male and 47% female, with a median age of 34. Domain expertise ranged from novices to individuals with over a decade of professional experience in law or healthcare. Statistical tests revealed no significant correlation between demographic attributes and participant dropout or incomplete sessions, supporting the assumption that the remaining dataset was representative of the original enrollment.

Each decision instance was annotated with metadata capturing prompt category, domain area, response content, identified hallucination severity, and user confidence level. The final consolidated dataset served as the foundation for advanced statistical modeling and hypothesis testing, described in subsequent sections. Descriptive statistics indicated variability in trust levels across domains, with legal tasks seeing slightly higher average confidence compared to medical tasks. This pattern aligns with the notion that certain professionals might hold a more conservative stance toward machine-generated medical advice due to the critical nature of patient outcomes. By contrast, legal practitioners may assume that textual references are easily verifiable through documented statutes, thus exhibiting a more moderate response to potential hallucinations.

Temporal patterns in user interactions emerged as well. Participants generally spent more time scrutinizing outputs in the middle of the session, while the beginning and end saw faster reading rates. This phenomenon might reflect an acclimation period where users learned how to navigate the interface before settling into a consistent review process. The subsequent data analysis accounts for these temporal nuances, ensuring that shifts in reliance or detection rates over the session are properly modeled.

4. Quantitative Analysis of Human Reliance

Statistical modeling began by examining bivariate relationships between hallucination presence and user reliance metrics. Reliance was operationalized through binary indicators (accept vs. reject) and continuous confidence ratings. Initial results showed a negative correlation between hallucination frequency and user acceptance rates, suggesting that users grew more skeptical when confronted with repeated inaccuracies. However, the magnitude of this effect varied by domain and user expertise level.

Variable	Coefficient	Standard Error	Significance Level
Hallucination Presence	-0.18	0.04	***
User Expertise (Moderate)	-0.12	0.05	**
Prompt Complexity	-0.08	0.03	*
Domain (Medical)	-0.25	0.06	***
Domain (Legal)	-0.14	0.05	**

Table 4. Logistic regression results predicting user acceptance of system outputs.

A logistic regression model was constructed to predict the probability of user acceptance based on hallucination presence, controlling for domain, difficulty level, participant expertise, and prompt complexity. Results indicated that hallucination presence reduced the likelihood of acceptance by an average of 18%, with confidence intervals indicating a robust effect across subgroups. Notably, users with moderate expertise were more vulnerable to hallucinated outputs, possibly due to partial familiarity with the domain that allowed them to miss subtle inaccuracies while trusting the model's overall authoritative tone.

Confidence ratings were assessed through linear mixed-effects models, with random intercepts for participants to account for intra-individual correlation. Hallucination presence, prompt domain, user expertise, and prompt difficulty served as fixed effects. Significant main effects emerged for both hallucination presence and prompt domain. The

Factor	Effect on Confidence	User Group Most Affected	Significance Level
Hallucination Presence	-0.22 (Mean Decrease)	Novices	***
Prompt Domain	-0.15 (Domain-Specific Variance)	Experts	**
Prior Exposure to Hallucinations	-0.10 (Cumulative Reduction)	General Users	*
Transparency Mechanisms	+0.05 (Minor Increase)	All Users	n.s.

Table 5. Effects of hallucination presence on user confidence ratings.

interaction between hallucination presence and expertise hinted that novices exhibited a more uniform decrease in confidence when they encountered any sign of factual mismatch, while experts showed a more selective pattern of reduced trust, rejecting certain inaccuracies outright but ignoring minor errors.

Time-to-decision analysis was performed to understand how hallucinations influence user deliberation speed. Survival analysis techniques, adapted for user interface data, treated each decision event as one in which the user either "accepted the system output" or "rejected it" at a certain time threshold. Observed hazard ratios showed that participants took longer to make acceptance or rejection decisions in the presence of hallucinations, indicating an additional cognitive burden to reconcile discrepancies or consult external knowledge. This finding was most pronounced when the content touched on complex or specialized topics.

Condition	Hazard Ratio	Effect on Decision Time	Interpretation
Hallucination Present	0.72	Increased deliberation time	Users hesitate before accepting/rejecting
High Complexity Prompt	0.81	Slower decision-making	Cognitive burden increases with complexity
Expert User Group	1.15	Faster rejection of hallucinations	Experts detect inconsistencies more quickly
Prior Hallucination Exposure	0.68	Extended response time	Users deliberate longer with repeated exposure

Table 6. Survival analysis results for time-to-decision in user interactions.

Multivariate analyses probed the extent to which repeated hallucinations cumulatively eroded trust. A cumulative measure of prior hallucination exposure was included in the regression models, capturing the number of instances where users had already encountered identified fabrications. Results showed a compounding effect: each additional hallucination faced previously increased the odds of rejection by around 6%, hinting at a memory-driven trust update process. However, this effect plateaued at higher exposure levels, suggesting that some users adopted a default stance of skepticism once they perceived multiple inaccuracies.

Further investigations extended to cluster analyses that grouped participants based on reliance patterns. Clustering algorithms such as k-means and hierarchical clustering segmented users into subsets based on their acceptance rates, confidence variance, and tolerance for minor hallucinations. One cluster demonstrated consistently high acceptance despite encountering hallucinations, reflecting a segment of users who defaulted to trusting system outputs, possibly due to a lack of time for manual verification or an assumption that minor errors would not be detrimental to overall task success. Another cluster took a conservative approach, rejecting outputs at the earliest sign of discrepancy, even when some content was verifiably correct.

Predictive modeling was also employed to see whether system outputs bearing explicit disclaimers about potential inaccuracies changed the reliance dynamic. Although disclaimers were not central to this study's design, a subset of prompts carried automated disclaimers inserted by the system in uncertain contexts. Analysis revealed a minor but statistically significant decrease in acceptance rates when disclaimers were present. Nonetheless, disclaimers alone did not mitigate the effect of repeated hallucinations on long-term trust erosion, pointing to the complexity of trust recalibration processes.

Cluster	**Acceptance Rate**	**Confidence Variance**	**Reliance Pattern**
High Trust Users	85%	Low	Accepts outputs despite hallucinations
Skeptical Users	40%	High	Rejects outputs at first sign of error
Moderate Reliance Users	60%	Medium	Evaluates outputs selectively
Adaptive Users	70%	Varies	Adjusts trust based on domain cues

Table 7. Clustering analysis of user reliance patterns based on acceptance and confidence levels.

Subgroup analyses by domain provided nuanced insights. In the medical domain, even a single detected hallucination triggered a sharp decline in trust, likely reflecting the high stakes associated with medical advice. Financial domain prompts, by contrast, showed more lenient responses, with participants sometimes disregarding minor inaccuracies in foreign exchange rates or minor cost estimates. The legal domain, as noted earlier, exhibited mid-level acceptance rates, with participants often cross-referencing textual clauses to confirm validity. Across all domains, user expertise intersected with domain-specific norms for fact-checking and risk assessment, shaping how individuals weighed the significance of discovered hallucinations [13,14].

Robustness checks validated these results. Sensitivity analyses omitted outliers, such as participants who exclusively accepted or rejected all system outputs. Additional models introduced lagged variables for prior confidence ratings, confirming that trust erosion was a dynamic process rather than a static effect tied to a single hallucination event. The overarching conclusion was that hallucinations, even when detected at moderate rates, significantly affected user reliance [15]. This impact was neither uniform across domains nor constant over time, reinforcing the need for a granular approach to measuring and interpreting the phenomenon.

5. Qualitative Observations and Discussion

Interview data and open-ended survey responses provided interpretative depth to the quantitative findings. Participants frequently recounted experiences of initial trust followed by disillusionment upon discovering an error. Some described a sense of betrayal, noting that the system's authoritative style gave them confidence before inaccuracies were uncovered. These sentiments align with theories of trust calibration that posit user trust oscillates in response to perceived reliability.

Observation of participants during eye-tracking sessions revealed that those who fixated on certain keywords—names, numbers, or specialized terminology—were more likely to detect hallucinations. The presence of numeric or factual details appears to function as an anchor point for scrutiny, possibly due to the ease of comparing such data against known references. Participants described employing heuristic checks, such as scanning for inconsistent numerical progressions or mismatched references to real-world locations. However, more subtle hallucinations slipped past these rudimentary checks, suggesting that reliance may persist when hallucinations involve conceptual or interpretive inaccuracies that lack easily verifiable markers [10].

The notion of partial reliance emerged prominently in interviews. Participants sometimes accepted segments of the generated text while disregarding or questioning suspicious portions. This selective use of the system's output complicated the binary metrics of accept vs. reject. A recurring theme was the perception that the LLM still provided baseline guidance or a workable template, even if certain details required correction. This attitude was especially prevalent among participants with moderate expertise, who felt equipped to correct minor mistakes but still benefited from the model's language structuring. However, the risk arises when hallucinations are not trivial, leading to profound misunderstandings.

Analysis of user attributions revealed that participants often rationalized hallucinations by attributing them to "small glitches" or "outdated training data," reinforcing a lenient stance toward the system. Others ascribed them to the model's inherent limitations, adopting a watchful skepticism. This differentiation in explanatory frameworks can shape how quickly participants reestablish trust after encountering an error. For instance, if a user believes that hallucinations stem from systemic issues related to data scope, they may be more vigilant in the future. Conversely, if they believe the system's intelligence outweighs these errors, they might maintain a relatively high level of trust even after repeated inaccuracies.

Group dynamics surfaced in cases where multiple participants collaborated on tasks, such as drafting a group document or collectively evaluating a legal scenario. Peer influence affected reliance: if one member questioned the system's accuracy, others often became more guarded. Conversely, if an authoritative group member endorsed the system's output, the entire group tended to adopt a more trusting stance. This phenomenon underscores the social dimension of reliance and suggests that hallucinations can have a magnified impact when group consensus is shaped by a single user's interpretation.

Some participants expressed frustration with the time required to validate the model's suggestions. They noted that while the system could generate comprehensive text quickly, verifying each detail to rule out hallucinations was time-consuming. The trade-off between speed and accuracy emerged as a pivotal factor in how reliance decisions were made, especially for professional tasks subject to tight deadlines. Users seeking fast drafting solutions might accept moderate hallucination risks, whereas those in high-stakes environments demanded thorough checking, which eroded any productivity gains.

Reflections on interface design surfaced repeatedly. Participants indicated that the presentation of generated text, such as how references or numeric data were highlighted, influenced how actively they looked for errors. Subliminal cues, including color-coding or interactive fact-check features, seemed to guide some participants toward more robust scrutiny. Nonetheless, the absence of universal interface standards means that user experiences varied, creating heterogeneous patterns of reliance.

Interpretation of the combined quantitative and qualitative data points to a complex interplay between the frequency of hallucinations, user expertise, task domain, and social context. Although hallucinations diminish trust, the rate at which reliance erodes depends on the user's initial orientation, domain norms, and the perceived risks of errors. Several participants continued to leverage LLM outputs even when they recognized recurrent inaccuracies, viewing them as placeholders or rough drafts that demanded manual oversight. Others adopted a zero-tolerance stance, discarding the system's outputs at the first sign of unreliability.

The emergent picture highlights the inadequacy of simplistic measurements that label an output as either correct or hallucinated. Users do not always treat these systems as sources of absolute truth; they adapt their trust levels dynamically based on experience, context, and external validations. Hallucinations act as focal points that accelerate or reinforce the process of trust recalibration, either undermining or validating preconceptions about the system's capabilities. By incorporating behavioral and perceptual dimensions, the study illuminates the real-world consequences of even intermittent hallucinations on user reliance.

6. Conclusion

Observations in this investigation underscore the intricate ways in which large language model hallucinations shape human reliance, highlighting the importance of a multidimensional evaluation framework. Even isolated inaccuracies can undermine confidence in system outputs, but the severity of this erosion depends on multiple factors, including user expertise, task risk, and the presence of prompt complexity. Quantitative analysis revealed that repeated hallucinations intensify skepticism, leading some participants to reject future outputs or invest additional time in scrutinizing each generated statement. Qualitative data expanded these findings by revealing the emotional and social dimensions of trust calibration, suggesting that users interpret hallucinations in terms of either minor glitches or systemic shortcomings.

Behavioral patterns observed in controlled experiments confirmed that hallucinations create pockets of doubt that may persist even in settings where the bulk of model output is accurate. The interplay between user vigilance, domain-specific norms, and the cognitive effort required to validate content emerged as central to reliance decisions. Overall, results emphasize that the challenge posed by hallucinations is not merely a matter of technical detection or correction but rather involves understanding the human side of trust formation and decision-making. Insights from this research may inform the design of future studies that integrate more granular user-level attributes with advanced detection methods, generating a comprehensive portrait of how hallucinations intersect with complex human behaviors.

References

1. Huschens, M.; Briesch, M.; Sobania, D.; Rothlauf, F. Do You Trust ChatGPT?–Perceived Credibility of Human and AI-Generated Content. *arXiv preprint arXiv:2309.02524* **2023**.
2. Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; et al. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In Proceedings of the NeurIPS, 2023.
3. Mehta, R.; Hoblitzell, A.; O'keefe, J.; Jang, H.; Varma, V. Halu-nlp at semeval-2024 task 6: Metacheckgpt-a multi-task hallucination detection using llm uncertainty and meta-models. In Proceedings of the Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), 2024, pp. 342–348.
4. Shen, X.; Chen, Z.; Backes, M.; Zhang, Y. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint arXiv:2304.08979* **2023**.
5. Niu, C.; Wu, Y.; Zhu, J.; Xu, S.; Shum, K.; Zhong, R.; Song, J.; Zhang, T. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *arXiv preprint arXiv:2401.00396* **2023**.
6. Bhaskaran, S.V. A Comparative Analysis of Batch, Real-Time, Stream Processing, and Lambda Architecture for Modern Analytics Workloads. *Applied Research in Artificial Intelligence and Cloud Computing* **2019**, *2*, 57–70.
7. McIntosh, T.R.; Liu, T.; Susnjak, T.; Watters, P.; Ng, A.; Halgamuge, M.N. A culturally sensitive test to evaluate nuanced gpt hallucination. *IEEE Transactions on Artificial Intelligence* **2023**.
8. Rawte, V.; Sheth, A.; Das, A. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922* **2023**.
9. Bhaskaran, S.V. Enterprise Data Architectures into a Unified and Secure Platform: Strategies for Redundancy Mitigation and Optimized Access Governance. *International Journal of Advanced Cybersecurity Systems, Technologies, and Applications* **2019**, *3*, 1–15.
10. Leiser, F.; Eckhardt, S.; Knaeble, M.; Maedche, A.; Schwabe, G.; Sunyaev, A. From ChatGPT to FactGPT: A participatory design study to mitigate the effects of large language model hallucinations on users. In *Proceedings of Mensch und Computer 2023*; 2023; pp. 81–90.
11. Jacob, C.; Kerrigan, P.; Bastos, M.T. The Chat-Chamber Effect: Trusting the AI Hallucination. *Big Data & Society, Forthcoming* **2023**.

12. Bhaskaran, S.V. Integrating Data Quality Services (DQS) in Big Data Ecosystems: Challenges, Best Practices, and Opportunities for Decision-Making. *Journal of Applied Big Data Analytics, Decision-Making, and Predictive Modelling Systems* **2020**, *4*, 1–12.
13. Guo, Z.; Xu, L.; Liu, J. Trustworthy Large Models in Vision: A Survey. *arXiv preprint arXiv:2311.09680* **2023**.
14. Dahan, S.; Bhambhoria, R.; Liang, D.; Zhu, X. Lawyers Should Not Trust AI: A call for an Open-source Legal Language Model. *Available at SSRN 4587092* **2023**.
15. Mehta, R.; Hoblitzell, A.; O'Keefe, J.; Jang, H.; Varma, V. MetaCheckGPT–A Multi-task Hallucination Detection Using LLM Uncertainty and Meta-models. *arXiv preprint arXiv:2404.06948* **2024**.