Human-centred (HCI) xAI - Hallucinations and user trust

Thomas Joyce * **Oisín Frizell** * **Andrew Jaffray** *
Edmund Phelan * **Ruairí Glackin** *

* *Immersive Software Engineering, University of Limerick, Castletroy,
Limerick (email: 23367857@studentmail.ul.ie,
23368276@studentmail.ul.ie, 23382163@studentmail.ul.ie,
23397179@studentmail.ul.ie, 23382732@studentmail.ul.ie)*

October 2025 | CS4437

Abstract: This study investigates hallucinations in Explainable AI (xAI) systems and their impact on user trust, combining theoretical analysis with empirical evaluation. Hallucinations; defined as high-confidence yet incorrect outputs, pose challenges for decision-making and can misalign user trust. We conducted a literature review to understand the underlying causes, including model stochasticity, data biases, and overgeneralization, and explored strategies for mitigating their effects through explanations and confidence indicators. Empirically, we evaluated three industry models using the HaluEval QA benchmark on 200 samples. Results show that Gemini-2.5-flash achieved 76% accuracy (152 correct), Meta-llama_Llama-3.1-8B 60% (120 correct), and DeepSeek V3.2 54% (108 correct), highlighting substantial variability in hallucination rates. Analysis indicates that hallucinations often arise from overconfident predictions and insufficient grounding in factual data, which can mislead users even when explanations are provided. These findings underscore the importance of transparent explanations and calibrated confidence measures in xAI systems to reduce trust misalignment. Overall, this work contributes to understanding the interplay between hallucinations, model performance, and user trust, offering guidance for designing more reliable and interpretable AI-assisted decision systems.

1. INTRODUCTION

Large Language Models (LLMs) have impressive fluency and ability to perform complex tasks but a critical challenge persists: hallucinations. Hallucinations occur when models generate output that appears coherent and convincing but is factually incorrect, fabricated, or logically inconsistent (Ji et al., 2023; Maynez et al., 2020; Kazemi et al., 2023). Numerous studies have documented this phenomenon, highlighting its prevalence and the risks it poses in high-stakes contexts with serious consequences, such as misdiagnoses in healthcare, fabricated citations in academic writing, and erroneous case references in legal documents [1], [2]. While some research has examined how human misconceptions or biases embedded in large-scale datasets contribute to model errors, these do not strictly constitute hallucinations. Similarly, trivial multiple-choice evaluation tasks may not capture the full impact of hallucinations, which are most misleading and consequential when arising from architectural limitations.

According to Hoang et al, hallucinations in LLMs can be divided into two primary sources: (1) Prompting-induced hallucinations, where ill-structured, unspecified, or misleading prompts cause inefficient outputs (Reynolds and McDonell, 2021; Zhou et al., 2022; Wei et al., 2022), and (2) Model-internal hallucinations, which caused by the model's architecture, pretraining data distribution, or inference behavior (Bang and Madotto, 2023; Chen et al., 2023; OpenAI, 2023a). Distinguishing between these two causes is essential for developing effective mitigation strategies.

Currently, there is no definitive method to prevent hallucinations. While RLHF has shown promise in improving model behavior, challenges remain in ensuring the accuracy and reliability of LLM outputs. To address this issue, we are conducting a literature review to study the interaction between LLMs and hallucinations. Our research involves benchmarking 3 different LLMs. We try to reproduce findings of ? HaluEval guys's analysis of hallucinations in large language models by using standardized hallucination evaluation benchmarks (Halluci-

Table 1. An example from Alpaca (Taori et al., 2023) gpt generate hallucinated con- tents (green) that cannot be verified by existing source.

User Query : Retrieve the oldest photo of a cat
(“ChatGPT”): There is no definitive answer to this question as “the oldest photo of a cat” is subjective and varies based on different sources and definitions. However, one can- didate for the oldest cat photo is the daguerreotype “Cat with Blue Ribbon,” taken by photographer Antoine Claudet in 1840. The photo depicts a cat sitting on a table, gazing directly at the camera, and wearing a blue ribbon tied around its neck. [...]

nationEval (Wu et al., 2023)). The conclusions drawn from this analysis aim to determine whether advancements in LLMs have mitigated the occurrence of hallucinations in previously problematic scenarios, and how this impacts user trust.

2. BENCHMARKING HALLUCINATIONS

2.1 Measuring Factuality and Harm: Standardized Benchmarks

Reliable evaluation of Large Language Models (LLMs) requires standardized, rigorous benchmarks that quantify different types of failure modes, including hallucinations, misinformation, and factual errors [1]–[3]. Hallucinations —outputs that are fluent but factually incorrect —pose serious challenges in high-stakes applications, making accurate assessment essential for both model development and user trust [1], [4], [5]. To systematically evaluate hallucination detection, we employed the HaluEval benchmark [6], which is specifically designed for QA, dialogue, and summarization tasks. HaluEval allows for precise measurement of a model’s ability to distinguish between factual and hallucinated content. We conducted a small-scale evaluation using a subset of 200 QA items balanced between correct and hallucinated answers. Each example consisted of an instruction, a question, an answer, and a corresponding judgment label. For example:

```
{
"knowledge": "Jonathan Stark (born April
3, 1971) is a former professional tennis
player from the United States. During
his career he won two Grand Slam doubles
titles (the 1994 French Open Men's
Doubles and the 1995 Wimbledon
Championships Mixed Doubles). He reached
the men's singles final at the French
Open in 1988, won the French Open men's
doubles title in 1984, and helped France
win the Davis Cup in 1991.",
"question": "Which tennis player won
more Grand Slam titles, Henri Leconte or
Jonathan Stark?",
"answer": "Henri Leconte won more Grand
Slam titles.",
"ground_truth": "Yes",
"judgement": "No"
}
```

This except identifies how models fail to parse truth for relatively complex logical/factual questions.

2.2 HaluEval Analysis

Each model was prompted with the instruction and asked to determine whether the provided answer was correct, yielding a binary Yes/No judgment. Model accuracy was then computed as the proportion of judgments aligning with the ground truth, following prior evaluation methodologies [6], [7].

Table 2. Hallucination detection results across different LLMs using HaluEval (QA subset, 200 samples). Accuracy is the fraction of judgments matching the ground truth.

Model	Correct	Incorrect	Accuracy
Gemini-3.5 Flash	152	48	0.76
Llama-3.1-8B-Instruct	120	80	0.60
DeepSeek V3.2 Exp	108	92	0.54

Gemini-3.5 Flash (76%) demonstrated strong factual awareness and reliably acted as a hallucination detection model, making it a suitable baseline for cross-model evaluation [6], [8]. Llama-3.1-8B-Instruct (60%) performed moderately, showing inconsistency in binary judgments and suggesting potential benefit from prompt tuning with explicit Yes/No constraints [9]. DeepSeek V3.2 Exp (54%) performed barely above random guessing, indicating weak hallucination detection capabilities and low reliability as a judge model [6], [10].

These findings underscore meaningful differences in hallucination detection ability across models and reinforce the importance of rigorous benchmark evaluations. By systematically comparing model judgments against human-verified ground truth, we can quantify reliability, assess the link between explainable AI (XAI), hallucination, and trust, and identify areas where architectural or prompting interventions may improve performance [1], [4], [11].

3. HALLUCINATIONS THROUGH A HUMAN–CENTRED LENS

Waiting on Ruairi's work...

3.1 User perception of hallucinations

3.2 Psychological and behavioral impacts

3.3 Case studies

4. XAI STRATEGIES FOR ADDRESSING HALLUCINATIONS

Hallucinations—false or fabricated answers generated by large language models (LLMs)—remain a major barrier to trustworthy AI. To tackle this challenge, xAI research has developed a range of detection and mitigation strategies, from model-internal mechanisms to user-facing interventions.

4.1 Detection and mitigation techniques

Recent work by Łajewska & Balog (2024) introduces a novel answerability detection strategy for conversational question answering. Their system operates in two sequential phases: first, it retrieves relevant passages from a candidate corpus as evidence; second, a sentence-level binary classifier determines whether the passages offer a valid answer. By gating responses with a boolean signal ("answerable" or "unanswerable"), the system avoids generating answers where none can be substantiated, measurably reducing hallucination rates and outperforming conventional LLM baselines in answerability prediction. This method sets a practical boundary for LLMs, directly preventing unsupported or fabricated replies and fostering user trust in AI outputs.

Complementing this, Su et al. (2024) propose an unsupervised, real-time framework (MIND) for hallucination detection using the internal states of LLMs during text generation. Unlike post-hoc fact-checking, their binary classification approach identifies hallucination as it occurs, leveraging internal model signals across models and domains. Evaluations against human annotation benchmarks show that MIND outperforms previous detection techniques, offering scalable and model-agnostic mitigation for xAI hallucination risks -Model-centric approaches (confidence estimation, uncertainty quantification)

-Interface-centric approaches (warnings, disclaimers, highlighting low-confidence outputs)

4.2 Human-centred design interventions

While technical advances are essential for limiting hallucinations, the way users interact with AI—through explanations, disclaimers, and transparent signals—plays a critical role in managing trust and user experience. Strategies such as confidence estimation, uncertainty visualization, and interactive explanations help users distinguish high-confidence from speculative outputs and calibrate their reliance accordingly. Designing UI elements that clearly indicate uncertain or unanswerable responses, informed by answerability or hallucination detection models, not only reduces the impact of errors but also supports informed decision-making.

4.3 Trade-offs and risks

Balancing safety and usability is key: overly conservative response gating may frustrate users, while insufficient mitigation increases the risk of misinformation. Overloading users with warnings can lead to desensitization, while effective xAI must remain intuitive and non-disruptive. Mitigating hallucinations vs preserving usability

Risks of reducing user autonomy through overcaution

5. TRUST CALIBRATION IN HUMAN–AI INTERACTION

5.1 Designing for calibrated trust/Measurement of trust erosion due to hallucinations

Calibrated trust refers to the alignment between a user's confidence in an AI system and the actual reliability or limitations of that system

[1]. Empirical studies have consistently shown that when AI systems generate hallucinations —outputs that are factually incorrect or misleading—user trust is significantly eroded [2]. For example, a 2024 controlled experiment demonstrated that user trust (as measured by TrustDiff scores) dropped rapidly after exposure to AI hallucinations, highlighting the precariousness of user confidence and the urgent need for accuracy in AI outputs [2]. In addition, research shows that overtrust (users relying too heavily on the AI) and undertrust (users disregarding accurate outputs) are common responses to poor calibration, both of which can lead to suboptimal or even harmful outcomes [1], [3]. Designing for calibrated trust thus involves not only increasing AI reliability but also implementing transparency tools, such as confidence indicators and explanatory warnings, that help users correctly interpret outputs [1], [3]. Studies suggest that adaptive cues—such as real-time notifications when users are over- or under-relying on the system—can improve trust calibration and task performance [3], [4]. Ultimately, the goal is to empower users to make informed decisions regarding when to accept, doubt, or verify the information provided by AI systems.

5.2 Evaluation frameworks

Evaluating trust calibration in human–AI interaction requires a multi-dimensional approach, combining quantitative task metrics with qualitative assessment of user perceptions. Researchers typically employ controlled user studies in which participants interact with AI systems under varying conditions of reliability and transparency. For instance, Sanderson et al. exposed participants to both accurate outputs and deliberate hallucinations, subsequently measuring shifts in user trust via validated survey instruments such as the TrustDiff and System Usability Scale [5]. Task performance metrics —like accuracy, completion rate, and error recovery—are analysed in tandem with subjective ratings of confidence and satisfaction [6], [7]. A key methodological development is the use of pre- and post-interaction surveys to gauge both anticipated and experienced trust levels [5], [6]. Wang et al. further refine this by integrating real-time monitoring of user reliance, noting moments of overtrust and undertrust during actual task execution [6]. Studies also indicate the value of behavioural analysis, such as tracking whether users seek external verification when confidence scores are low [5], [6]. Comparative frameworks analyse results across groups exposed to varying transparency cues (e.g., confidence indicators, warnings), enabling researchers to isolate the effect of specific design interventions on trust calibration [7]. Ultimately, these evaluation frameworks reveal not only the impact of hallucinations on user trust but also how interface design and feedback mechanisms can mitigate or exacerbate trust erosion [5], [7]. By triangulating empirical data from task outcomes, surveys, and behavioral traces, researchers gain a robust understanding of how to promote calibrated trust in complex xAI systems [6], [7].

5.3 Ethical considerations

Evaluating the increasing deployment of explainable AI systems raises complex ethical questions regarding responsibility and accountability, particularly when AI-generated hallucinations mislead users [8], [9]. When an AI system presents inaccurate information, determining responsibility becomes a challenge: should liability rest with the system's designers, its operators, or the end-users themselves [10], [11]? Studies highlight that users often place implicit trust in AI output, expecting systems to provide reliable information; as such, misleading hallucinations can result in tangible harm—especially in sensitive domains such as healthcare or finance [8], [10].

Transparency is a key strategy for mitigating these risks [9], [10]. By proactively communicating uncertainties, limitations, and confidence scores, developers can help users understand when and why the system's outputs may be unreliable [2]. However, there is a tension between transparency and liability: providing greater warning or disclosure may reduce the risk of misuse, but it can also increase the burden of responsibility on solution providers [9], [11]. Furthermore, excessive emphasis on uncertainty may undermine user confidence and adoption—even in situations where the system's outputs are generally reliable [9].

To address these dilemmas, best practices recommend clear documentation, user education, and ongoing monitoring of AI outputs [9], [10]. Automated error reporting and feedback loops can alert developers when hallucinations have the potential to mislead, enabling rapid mitigation of errors [10]. Crucially, ethical frameworks suggest that responsibility should be shared: designers and operators must adopt robust technical safeguards, while users are empowered to critically assess outputs—guided by transparent cues and explanatory warnings [9], [11]. The collaborative management of trust calibration thus becomes central to minimising risk and ensuring responsible AI deployment [8], [9].

6. OPEN CHALLENGES AND RESEARCH GAPS

6.1 How to measure the degree of harmfulness of a hallucination

As Artificial intelligence becomes increasingly common in everyday life the challenge of addressing hallucinations remains a critical issue. One major research gap is how to effectively measure the harmfulness of and hallucinations. Current studies classify hallucinations as a variety of distorted information types ranging from logic errors to factual inaccuracies. Each hallucination carries a degree of potential harm depending on the context. For example in healthcare a hallucinated diagnosis could lead to serious patient harm while in other domains consequences might hurt reputations or finances. The complexity of evaluating harm arises from multifaceted criteria that include not only the factual inaccuracy but also the context and severity of outcomes. (1)Researchers have developed taxonomies of error types specific to AI-generated content and have begun creating coding schemes validated through empirical data such as error reports from users interacting with large language models to better characterize and quantify these harms. (2) However these efforts highlight the difficulty of standardizing such evaluations as the current taxonomies are often subjective or limited to isolated domains. This puts urgent need for generalized frameworks that can categorize and measure hallucination harm across contexts.

6.2 Cross-cultural differences in trust calibration

Trust in AI isn't universal either, it is deeply influenced by cultural values and past experience with technology. Studies show that strategies for building trust in one region might fail in another. Users in particular cultures might expect a more detailed explanation or show greater skepticism towards automated decisions(1). Disparities in IT knowledge, access to technology and prior AI exposures further shape how trust is formed. In order to design inclusive AI systems researchers advocate for different approaches including anthropology, linguistics and psychology.

6.3 Long-term trust dynamics in everyday use

Most research in AI trust docs on short-term interactions in controlled settings, but real world AI use unfolds over weeks, months or years. Repeated experiences of accurate results can lead to a steadily build in confidence however it can only take 1 hallucination to break trust. (2)Emerging concepts like metacognitive sensitivity in which users learn to recognize the limits of AI reliability and adjust their reliance accordingly highlight the need for dynamic trust calibration tools that continuously educate users and adapt to their changing experiences and expectations.

6.4 Standardising evaluation of hallucination impacts

Continuing there is still a research gap when it comes to the standardization of evaluation metrics for hallucination impacts. Despite the many benchmarks for hallucination detection and assessment, they often differ on fundamental definitions i.e what counts as a hallucination and how do we quantify its severity and damage. THe lack of consistency prevents fair comparisons across models and stops efforts to identify which systems minimize harmful errors

See Able (1956), Able et al. (1954), Keohane (1958), and Powers (1985).

7. CONCLUSION

TBC

REFERENCES

Able, B. (1956). Nucleic acid content of microscope. *Nature*, *135*, 7–9.

Able, B., Tagg, R., & Rush, M. (1954). Enzyme-catalyzed cellular transanimations. In A. Round (Ed.), *Advances in Enzymology* (3rd ed., Vol. 2, pp. 125–247). Academic Press.

Keohane, R. (1958). *Power and Interdependence World Politics in Transitions.* Little, Brown & Co.

Powers, T. (1985). Is there a way out?. *Harpers*, 35–47.

Appendix A. DECLARATIONS

7.1 AI Declaration

Parts of this manuscript were generated or assisted by large language models, including text summarization, editing, and organization. The authors have verified the factual accuracy and intellectual content of all AI-generated material, and any errors or misrepresentations remain the responsibility of the authors. The use of AI tools does not replace critical evaluation, scholarly judgment, or original analysis.

7.2 Equal Work

page that is signed by all group members attesting to the satisfaction that all members contributed equally to the creation of the report and that it is your own work.

Signed : Thomas Joyce, 09/10/2025
Signed : rg, 09/10/2025
Signed : ep, 09/10/2025
Signed : of, 09/10/2025
Signed : andrew, 09/10/2025