# 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 Large Language Model (LLM) systems and their impact on user trust. 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 overgeneralisation, and explored strategies for mitigating their effects through explanations and confidence indicators.

We evaluated three models using the HaluEval QA (Question Answering) [1] 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 across various LLMs. These differences likely stem from variations in training data quality, alignment techniques, and model size, which strongly influence factual reliability and reasoning consistency.

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.

## 1. INTRODUCTION

Large Language Models 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 [2], [3], [4]. Numerous studies have documented this phenomenon, highlighting both its prevalence and the risks it poses in high-stakes contexts such as healthcare misdiagnoses [5], fabricated citations in academic writing [6], and erroneous legal case references. These cases underscore that hallucinations are not merely linguistic artifacts but can lead to significant real-world consequences.

According to [7], hallucinations in LLMs can be divided into two primary sources:
(1) Prompting-induced hallucinations, where ill-structured, unspecified, or misleading prompts cause inefficient outputs [8], [9], [10]
(2) Model-internal hallucinations, which caused by the model's architecture, pretraining data distribution, or inference behavior [11], [12], [13].

Distinguishing between these two causes is essential for developing effective mitigation strategies.

Despite advances such as Reinforcement Learning from Human Feedback (RLHF), no definitive solution exists for eliminating hallucinations. Ensuring factual reliability and epistemic trustworthiness in LLMs therefore remains an open challenge. To address this gap, the present study conducts a targeted literature review and empirical benchmarking of three contemporary LLMs. Specifically, we replicate and extend the HaluEval [1] analysis using standardised hallucination evaluation benchmarks to assess whether recent model improvements have reduced hallucination frequency. The findings aim to clarify whether progress in model design has mitigated prior failure modes and to evaluate the implications for user trust and model reliability.

**User Query : Where is the Immersive Software Engineering building in the University of Limerick**

("ChatGPT"): The Immersive Software Engineering (ISE) programme at the University of Limerick (UL) is primarily based in the Science & Engineering Faculty Office, located on the 1st Floor of the Lonsdale Building on the UL campus in Castletroy, Limerick, Ireland. This building houses the Department of Computer Science and Information Systems, which oversees the ISE programme [...]

Table 1. An example of hallucinated content [14].

## 2. BENCHMARKING HALLUCINATIONS

### *2.1 Measuring Factuality and Harm: Standardised Benchmarks*

Reliable evaluation of Large Language Models requires standardised, rigorous benchmarks that quantify different types of failure modes, including hallucinations, misinformation, and factual errors [15], [16]. To systematically evaluate hallucination detection, we employed the HaluEval benchmark [1], which is specifically designed for dialogue, and summarisation 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 excerpt identifies how models fail to parse truth for 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 [1], [17].

Table 1. 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 [1], [18]. 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 [19]. DeepSeek V3.2 Exp (54%) performed barely above random guessing, indicating weak hallucination detection capabilities and low reliability as a judge model [1], [20].

These findings underscore substantial variability in hallucination-detection ability across contemporary LLMs and highlight the need for standardised, reproducible benchmark evaluations. Gemini's higher performance likely reflects architectural and alignment improvements such as advanced reinforcement learning from human feedback and enhanced factual consistency mechanisms which mitigate hallucination generation and recognition. In contrast, the DeepSeek model, while valuable for transparency and research reproducibility, appears to lack comparable alignment tuning, resulting in higher error rates. This disparity illustrates the evolving trade-off between openness and reliability in the current LLM landscape.

## 3. HALLUCINATIONS THROUGH A HUMAN–CENTRED LENS

### *3.1 User perception of hallucinations*

From a human-centred perspective, the recognition of hallucinations depends less on technical expertise and more on perceived coherence, plausibility, and contextual fit. According to [21], among millions of user reviews of AI mobile applications, only around 1.75% explicitly mentioned hallucination-like errors, even though such errors were likely more common. Users rarely used the technical term "hallucination", instead describing the phenomenon in everyday language such as "it made this up," "AI lies", or "nonsense answer." This suggests that non-experts detect hallucinations heuristically, typically when an output violates basic common knowledge, logical reasoning, or conversational norms.

User perception of hallucinations is reactive rather than proactive. Users are more likely to identify blatant contradictions than to question plausible but incorrect claims. Because language models present information with linguistic confidence, users often equate fluency with accuracy, leading to misplaced trust in generated outputs.

### *3.2 Psychological and behavioral impacts - HCI*

The psychological effects of hallucinations can be described along a continuum of trust calibration, where user experiences shape emotional and behavioral responses.

In the early stages of interaction, users often display overtrust, placing excessive confidence in AI systems due to their coherent and authoritative communication style. When hallucinations occur, this trust can be violated, resulting in frustration or feelings of betrayal. According to [21], hallucination-related reviews were associated with strong negative sentiment (mean VADER (Valence Aware Dictionary and sEntiment Reasoner) score = $-0.65$) and lower app ratings (approximately 1.8/5) compared with general AI reviews (around 3.9/5), demonstrating the erosion of user trust once errors become apparent.

With repeated exposure, users may develop undertrust, becoming overly skeptical and distrusting even accurate outputs. Bucina et al. [22] relate this pattern to the lack of design mechanisms that promote careful reflection. Their discussion of "cognitive forcing functions" as interventions highlights how small design frictions could encourage users to pause and verify information before acting on it. Without such mechanisms, users alternate between inattentive overreliance and blanket rejection of AI outputs.

This cycle reflects a broader human tendency to offload cognitive effort to automation until errors expose system limitations. Addressing this requires transparent confidence indicators and interaction designs that re-engage users' critical attention when accuracy matters most.

### *3.3 Case studies*

Conversational AI systems such as ChatGPT and Gemini provide practical contexts for examining hallucinations and their impact on user trust. In their large-scale review analysis, [21] identified common hallucination types in such systems, including factual inaccuracies ("gave the wrong historical date"), irrelevant or repetitive responses ("it kept repeating words"), and inconsistent personas ("it suddenly changed tone"). These user accounts show that hallucination perception varies by task type: factual domains elicit scrutiny of correctness, while conversational or creative tasks highlight coherence and tone.

## 4. XAI STRATEGIES FOR ADDRESSING HALLUCINATIONS

Recent research has proposed a wide range of detection and mitigation strategies for hallucinations. These strategies range from internal-mechanisms in models to interventions that target the users of LLMs.

### *4.1 Detection and mitigation techniques*

Recent work by [15] introduces a new strategy for answerability detection for conversational question answering. The system they have developed has two phases, it retrieves the relevant passages that might answer the question, it then uses a classifier to check if these passages truly contain an answer. If not, the system labels the question as "unanswerable" and avoids giving a potentially wrong answer. This boolean value allows the AI to only answer a question when it is sure that it has a correct answer. This technique performs better than standard baselines for LLMs in answerability prediction. This reduces the hallucination rates of LLMs considerably. However, this is unfeasible to implement in production in practicality.

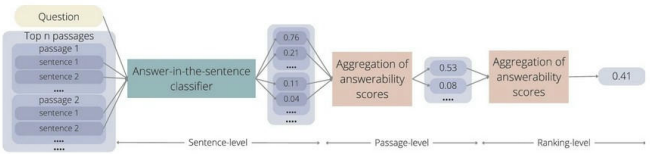


Fig. 1. Łajewska & Balog (2024). Overview of answerability detection approach.

Another method that complements this, [23] propose a real-time framework that is unsupervised (MIND) for detecting hallucinations. This is done using the internal state of LLMs during text generation. It uses a binary classifier to spot signs of hallucination during text generation using internal

model signals across different models and domains. When tested against human-annotated benchmarks, this approach outperformed previous detection techniques. This method can be scaled and is model-agnostic, so it can potentially be used in many AI systems.

### *4.2 Human-centred design interventions*

While technical advancements are essential for reducing hallucinations in LLMs, the way users interact with AI systems is just as important for managing trust and the user experience. This can be done through the use of clear explanations from AI systems justifying their answers, especially when the topic is complex. This helps users see how the system reached its conclusion, and encourages their own critical thinking and thought process, reducing the overall 'blind' trust that is often expected from LLMs.

Visual cues such as confidence bars or probability scores help users judge quickly how reliable an answer might be. A low confidence bar might encourage a user to double-check a fact before relying on an LLM to tell them. Other methods such as warnings or disclaimers may assist a user in understanding the limitations of a particular system. Prompting a user with "This answer may not be accurate" is a method of discouraging an overreliance on an AI's outputs. Using different types of explanations, such as sources or counter-examples can also be beneficial to users with different needs.

### *4.3 Trade-offs and risks*

Designing an explainable AI system that is resilient against hallucinations involves important trade-offs. While using strict detection and warnings can help prevent false information from reaching a user, this can also lead to the model being overly conservative in its responses and reduce the overall usefulness of the AI. If an AI system is too cautious, it may refuse to answer questions or bombard a user with warnings after every prompt. This can be frustrating to a user and limit the experience of interacting with the AI.

On the other hand, if the system is too relaxed and lacks mitigation, hallucinations may go unnoticed and potentially spread misinformation which can be harmful to a user. Warnings to a user must be deliberate and not appear too often. A user could become desensitised to the amount of warnings they receive and disregard them altogether. The ultimate goal is to help users trust the AI system without blindly accepting everything it says. It should be used as a tool, rather than an oracle. This means that the AI must be transparent about uncertainty, but not overwhelm users with alerts or blocking answers unnecessarily.

## 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 [15]. 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 [15]. 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 [15]. 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 [18]. 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. By exposing participants to both accurate outputs and deliberate hallucinations, shifts in user trust can be subsequently validated via survey instruments such as the Trust-Diff and System Usability Scale [24].

Task performance metrics like accuracy, completion rate, and error recovery are analysed in tandem with subjective ratings of confidence and satisfaction [19], [25]. A key methodological development is the use of pre- and post-interaction surveys to gauge both anticipated and experienced trust levels [24], [19]. [24] further refine this by integrating real-time monitoring of user reliance, noting moments of overtrust and undertrust during actual task execution [19]. Studies also indicate the value of behavioural analysis, such as tracking whether users seek external verification when confidence scores are low [24], [19].

Comparative frameworks analyse results across groups exposed to varying transparency cues (e.g., confidence indicators, warnings), enabling re-

searchers to isolate the effect of specific design interventions on trust calibration [25]. 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 [24], [25]. 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 [19], [25].

### *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 [20], [26]. 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 [27], [21]? 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 [20], [27].

Transparency is a key strategy for mitigating these risks [26], [20]. By proactively communicating uncertainties, limitations, and confidence scores, developers can help users understand when and why the system's outputs may be unreliable [16]. 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 [26], [21]. Furthermore, excessive emphasis on uncertainty may undermine user confidence and adoption even in situations where the system's outputs are generally reliable [26].

To address these dilemmas, best practices recommend clear documentation, user education, and ongoing monitoring of AI outputs [26], [20]. Automated error reporting and feedback loops can alert developers when hallucinations have the potential to mislead, enabling rapid mitigation of errors [20]. 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 [26], [21]. The collaborative management of trust calibration thus becomes central to minimising risk and ensuring responsible AI deployment [20], [26].

## 6. OPEN CHALLENGES AND RESEARCH GAPS

### *6.1 How to measure the degree of harmfulness of a hallucination*

The complexity of evaluating harm arises from multifaceted criteria that include not only the factual inaccuracy but also the context and severity of outcomes. 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 characterise and quantify these harms [28]. However these efforts highlight the difficulty of standardising such evaluations as the current taxonomies are often subjective or limited to isolated domains. This puts urgent need for generalised frameworks that can categorise and measure hallucination harm across contexts [29].

### *6.2 Hallucination's Technical Cause*

Hallucination in large language models originates from both architectural constraints and contextual limitations inherent to transformer-based generation. Architecturally, LLMs such as GPT variants are trained via next-token prediction using maximum likelihood estimation, which optimizes for local coherence rather than factual grounding. This objective causes models to privilege probabilistic fluency over epistemic accuracy, generating tokens that best fit preceding text even when semantically false. The self-attention mechanism amplifies this behavior by weighting syntactic relevance over factual consistency, lacking any explicit grounding in external truth signals or world models. \ Moreover, due to the finite context window and absence of persistent memory, models lose earlier factual constraints as generation proceeds, producing drifted continuations that sound credible but diverge from truth. Contextually, hallucination often emerges from underspecified or open-ended prompts, where the model must extrapolate beyond its training distribution or synthesize unseen entities, as analyzed in [15]. \ Compounding this, internal confidence signals are frequently miscalibrated: token probabilities do not reliably reflect truth likelihood, leading models to assert fabricated content with high certainty, a dynamic examined in [18]. Finally, from an epistemic standpoint, hallucination reflects an absence of grounded self-monitoring, LLMs lack mechanisms to assess whether a generated statement is verifiable or answerable, echoing the broader ethical and cognitive parallels discussed in [16].

### *6.3 Long-term trust dynamics in everyday use*

Most research in AI trust focuses 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 [29]. Emerging concepts like metacognitive sensitivity in which users learn to recognise 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 [29].

### *6.4 Standardising evaluation of hallucination impacts*

There is still a research gap when it comes to the standardisation 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 minimise harmful errors.

See [15]

## 7. CONCLUSION

This study confirms that hallucinations remain a critical challenge for Large Language Models, significantly affecting model reliability and user trust. Our empirical benchmarking using the HaluEval QA subset demonstrated substantial variability in hallucination detection ability across contemporary models, with Gemini-3.5 Flash showing higher factual consistency than Meta-Llama-3.1-8B and DeepSeek V3.2. These findings underscore the need for continuous, standardised evaluation to quantify reliability [1].

The analysis of user perception reveals that non-experts typically identify hallucinations not through technical means, but heuristically, describing the failure as the AI "lying" or providing "nonsense," which leads to strong negative emotional responses and trust erosion [21]. This human-centric view is crucial because LLM confidence and fluency often lead to overtrust, where users mistake coherence for correctness.

To counter this, xAI strategies must focus on promoting calibrated trust. Technical interventions, such as answerability detection by [15] and real-time internal state monitoring by [23], are vital for reducing the generation of factually unsupported output. These must be complemented by human-centred design elements, including transparent confidence indicators and contextual warnings, to help users interpret model outputs critically [19], [24].

Despite progress, significant open challenges remain. The document highlights the difficulty in standardising the measurement of hallucination harmfulness across diverse contexts and the need to investigate cross-cultural differences in trust calibration. Furthermore, current evaluation benchmarks, such as the one used in this study, primarily measure recognition or recall and often fail to simulate open-ended, real-world tasks where hallucinations cause real harm (e.g., legal advice, medical reasoning, research synthesis). 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.

Ultimately, mitigating the risk of hallucinations and ensuring ethical deployment requires a shared responsibility model [26]. Developers must implement robust technical safeguards and transparency mechanisms, while users must be educated and empowered to critically assess AI outputs. Future research should prioritise investigating long-term trust dynamics in everyday use and developing standardised, ecologically valid frameworks to measure the true societal impact of LLM hallucinations.

REFERENCES


[1] J. W. et al., "HaluEval: A Universal Evaluation Benchmark for Hallucination in Large Language Models," in *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[2] Z. J. et al., "Survey of Hallucination in Large Language Models," *ACM Computing Surveys*, 2023.

[3] J. M. et al., "Faithful to the Original: Factuality and Attribution in Abstractive Summarization," in *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[4] S. M. K. et al., "Hallucinations in Large Language Models: A Survey," *arXiv preprint*, 2023.

[5] H. N. et al., "Capabilities of GPT-4 in Medical and Clinical Domains," *NPJ Digital Medicine*, 2023.

[6] G. Zuccon, "ChatGPT and Scholarly Hallucinations," *Nature*, 2023.

[7] Q. H. et al., "A Classification of Hallucinations in Large Language Models," *ArXiv Preprint*, 2024.

[8] L. Reynolds and J. McDonell, "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm," in *NeurIPS Workshop on Distribution Shifts*, 2021.

[9] S. Z. et al., "Controllable Text Generation with Language Models," *ArXiv Preprint*, 2022.

[10] J. W. et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[11] Y. Bang and E. Madotto, "A Categorization of Hallucination Types in LLMs," *ArXiv Preprint*, 2023.

[12] T. C. et al., "Survey on Hallucination in Large Language Models," *ArXiv Preprint*, 2023.

[13] OpenAI, "GPT-4 Technical Report." 2023.

[14] (OpenAI) gpt-5 model, "gpt5 Hallucination - https://imgur.com/NYKDNOQ." 2025.

[15] W. Łajewska and K. Balog, "Towards Reliable and Factual Response Generation: Detecting Unanswerable Questions in Information-Seeking Conversations," *Lecture Notes in Computer Science*. Springer, Cham, pp. 317–328, 2024.

[16] AI Hallucinations? What About Human Hallucination?! Addressing Human Imperfection Is Needed for an Ethical AI, "AI Hallucinations? What About Human Hallucination?! Address-

ing Human Imperfection Is Needed for an Ethical AI," *International Journal of Interactive Multimedia and Artificial Intelligence (IJIMAI)*, vol. 9, no. 2, pp. 8–18, 2025.

[17] H. V. et al., "Explanations Can Reduce Overreliance on AI Systems During Decision-Making," *Proc. ACM Hum.-Comput. Interact. (CSCW)*, vol. 7, no. CSCW1, pp. 1–38, 2023.

[18] Understanding the Effects of Miscalibrated AI Confidence on Human Trust, "Understanding the Effects of Miscalibrated AI Confidence on Human Trust." 2024.

[19] V. L. et al., "Effect of Confidence and Explanation on Accuracy and Trust Calibration in Human-AI Collaboration." 2020.

[20] Fake it till you make it? AI hallucinations and ethical responsibility in medicine, "Fake it till you make it? AI hallucinations and ethical responsibility in medicine." 2025.

[21] R. M. et al., ""My AI is Lying to Me": User-reported LLM hallucinations in AI mobile apps reviews," *Sci Rep.*, vol. 15, p. 30397, 2025, doi: 10.1038/s41598-025-15416-8.

[22] Z. Bucinca, M. B. Malaya, and K. Z. Gajos, "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–21, 2021, doi: 10.1145/3449287.

[23] W. S. et al., "Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models." 2024.

[24] D. W. et al., "Measuring and Understanding Trust Calibrations for Explainable AI," in *Proc. ACM Conf.*, 2023.

[25] S. Sun A. Kulkarni and B. Y. Lim, ""Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction," in *CHI '23 (Best Paper HM)*, 2023.

[26] Transparency and accountability in AI systems, "Transparency and accountability in AI systems," *Frontiers in Human Dynamics*, 2024.

[27] AI Hallucinations: When Creation Comes at a Cost, Who Pays?, "AI Hallucinations: When Creation Comes at a Cost, Who Pays?." 2025.

[28] M. N. et al., "Explainable Recommendations and Calibrated Trust: Two Systematic User Errors," *Computer (IEEE)*, vol. 54, no. 10, pp. 28–37, 2021, doi: 10.1109/MC.2021.3076131.

[29] E. M. B. et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," in *FAccT '21*, 2021. doi: 10.1145/3442188.3445922.

## Appendix A. DECLARATIONS

### *7.1 AI Declaration of Use*

Parts of this manuscript were spell-checked or assisted by large language models (https://chatgpt.com/), including text summarisation, editing, and organisation. 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*

We are satisfied that all members contributed equally to the creation of the report and that it is our own work.
Signed : Edmund Phelan, 09/10/2025
Signed : Thomas Joyce, 09/10/2025
Signed : Ruairí Glackin, 09/10/2025
Signed : Oisín Frizell, 09/10/2025
Signed : Andrew Jaffray, 09/10/2025