ISSN–2278-5655

AMIERJ Aarhat Multidisciplinary International Education Research Journal

Volume–XV, Special Issues– I(b) Jan – Feb, 2026

OPEN ACCESS

Original Research Article

# A COMPARATIVE REVIEW OF HALLUCINATIONS IN LARGE LANGUAGE MODELS AND HUMAN PERCEPTIONS OF BIAS

*** Muzammil Mehboob Khan**

*Students, K S D's Model College (Empowered Autonomous), Khambalpada Rd, Thakurli, Dombivli East, Maharashtra*

**Abstract:**

*Large Language Models (LLMs) have become integral to a wide range of applications, raising concerns about their tendency to generate hallucinated content and exhibit biases inherited from training data. While prior research has examined hallucination behavior across different AI models, less attention has been given to how these limitations align with human perceptions of bias and trust.*

*This paper presents a comparative review of existing research on hallucinations in contemporary LLMs, synthesizing findings across multiple studies to identify common trends, evaluation approaches, and reported limitations. In parallel, a human perception study examines how users interpret and judge bias, reliability, and trustworthiness in AI-generated outputs. Participants provide subjective assessments of perceived bias and confidence in model responses, enabling comparison with conclusions drawn in prior technical literature.*

*The findings reveal a clear divergence between empirically reported hallucination behavior and user perception. Models identified as having lower hallucination tendencies are not consistently perceived as less biased or more trustworthy. Instead, fluent and confident responses often lead to higher perceived reliability, regardless of documented limitations. This highlights a disconnect between technical evaluation and human judgment.*

*This study emphasizes integrating human-centered perspectives into LLM evaluation and underscores the need for transparency, clearer communication of limitations, and trust-aware deployment.*

**Keywords:** *Large Language Models, Hallucinations, Human Perception, Automation Bias, AI.*

## Introduction:

The contemporary information ecosystem is undergoing a seismic shift, moving from a paradigm of "information filtering" (search engines ranking human-authored content) to "information generation" (LLMs synthesizing answers). This transition fundamentally alters the nature of digital truth. Unlike a search engine, which directs a user to a source, an LLM acts as an oracle, generating a plausible-sounding answer token by token based on probabilistic weightings.

The implications of this shift are profound. As generative systems like GPT-4o, Claude 3.5, and DeepSeek-R1 become the primary interface for knowledge retrieval, the risk is no longer just finding "bad sources," but encountering "hallucinations"—authoritative fabrication generated by the tool itself.

This report explores this mismatch through two parallel lenses: the technical reality of model hallucinations (how often they occur and why) and the psychological reality of human perception (why we believe them). By integrating 2024-2025 benchmarking data with behavioral analysis, we argue that the core danger of LLMs is not just their inaccuracy, but the "epistemic alignment" between their confident tone and human cognitive biases.

**Literature Review:**

**1. Theoretical Framework: Stochastic Parrots and the Taxonomy of Error**

To rigorously evaluate the reliability of LLMs, one must first disaggregate the concept of hallucination from simple error. A hallucination is not merely a mistake; it is a "fluent failure"—an error that mimics the structural properties of truth.

**The "Stochastic Parrot" Hypothesis and Beyond**

The critique of LLMs as "stochastic parrots"—a metaphor introduced by Bender et al. (2021)—argues that these models are statistical mimics. They learn the probability distribution of word sequences but lack an underlying model of the world or "communicative intent." When an LLM asserts a fact, it is not "telling the truth"; it is simply predicting that the word "Paris" is the most likely token to follow "The capital of France is." Even sophisticated models that utilize "Chain of Thought" (CoT) processing are ultimately predicting the next step in a reasoning chain based on training patterns, not verifiable logic. This makes them prone to "confabulation," where they invent facts to maintain the fluency of the narrative.

**A Taxonomy of Hallucination: Intrinsic vs. Extrinsic** Effective benchmarking requires a precise taxonomy. The 2025 "HalluLens" and "RealtimeQA" frameworks divide hallucinations into two primary categories:

- **Intrinsic Hallucinations:** Occur when the generated output contradicts the source content provided in the prompt.
  - *Example:* In a summarization task, the source text states "Company X revenue fell by 5%," but the LLM summary says "Company X revenue rose by 5%."
  - *Mechanism:* This represents a failure in the attention mechanism's ability to attend to the relevant tokens in the context window.
- **Extrinsic Hallucinations:** Occur when the generated output cannot be verified from the source input and is factually incorrect regarding the real world.
  - *Example:* When asked to write a biography of a minor historical figure, the model invents a birth date or middle name that sounds plausible but is false.
  - *Mechanism:* This is a failure of "parametric knowledge." The model is filling gaps in its training distribution with statistically likely—but factually wrong—information.

**2. The 2024-2025 Benchmark Landscape**

The evaluation of LLM reliability has matured from anecdotal "vibes-based" assessment to rigorous, automated benchmarking.

**The "Accuracy Paradox" in Reasoning Models**

One of the most counterintuitive findings of 2025 is the "Accuracy Paradox": models optimized for advanced reasoning often perform *worse* on simple factual recall than less capable models.

- *OpenAI's SimpleQA and the o-Series Regression:* OpenAI's "SimpleQA" benchmark, designed to measure short-form parametric factuality, revealed a startling regression. The o4-mini model, despite being optimized for efficiency and reasoning, demonstrated a hallucination rate significantly higher than the older GPT-4-Turbo on obscure facts.
- *PersonQA Performance:* On the PersonQA benchmark, the older o1-preview model outperformed the newer, "smarter" o1 model.
- *Mechanism of Failure:* The leading hypothesis is the "Reward for Confidence." In Reinforcement Learning from Human Feedback (RLHF),

models are rewarded for producing definitive answers. "Reasoning" models essentially "over-think" simple queries, generating complex but wrong justifications.

**The DeepSeek Divergence: R1 vs. V3** Data from the Vectara Hallucination Leaderboard corroborates this paradox:

- *DeepSeek-V3:* A standard model, exhibited a hallucination rate of roughly 1.8% on summarization tasks.
- *DeepSeek-R1:* The "reasoning" variant exhibited a hallucination rate jumping to nearly **14.3%** in certain contexts. This suggests that the "reasoning" process introduces generative noise.

**Table1: The Vectara Leaderboard: The State of the Art (2025) As of early 2025, the Vectara Leaderboard provides one of the most comprehensive comparisons of model faithfulness**.

| **Rank** | **Model** | **Hallucination Rate** | **Answer Rate** | **Average Summary Length** |
|---|---|---|---|---|
| 1 | Gemini-2.0-Flash-001 | 0.7% | N/A | ~60 words |
| 2 | Gemini-2.0-Pro-Exp | 0.8% | N/A | ~65 words |
| 3 | GPT-4o | 1.5% | N/A | ~86 words |
| 4 | GPT-4-Turbo | 1.7% | N/A | ~86 words |
| 5 | GPT-3.5-Turbo | 1.9% | N/A | ~84 words |
| ... | ... | ... | ... | ... |
| Low | DeepSeek-R1 | 14.3% | High | ~93 words |
| Low | Microsoft Copilot Pro | 27% | N/A | N/A |

*Source: Vectara (2025); Xu et al. (2025).*

## 3. Domain-Specific Hallucinations: High Stakes and Hard Failures

While general benchmarks provide a baseline, the impact of hallucinations is most acute in specialized domains.

- **The Legal Sector:** The legal domain is vulnerable to citation fabrication. Stanford HAI researchers found that general-purpose LLMs hallucinate on legal queries up to **17-30%** of the time. This occurs because models learn the *style* of a citation perfectly, even when the numbers point to non-existent cases.
- **Healthcare:** In healthcare, "sycophantic" models can be life-threatening. A deployment at Harris Health System achieved a positive predictive value of only roughly 50% for sepsis alerts. A meta-analysis of oncology questions found a pooled hallucination rate of **12.6%**.
- **Software Supply Chain:** "Package Hallucination" has emerged as a security threat.

Models may hallucinate plausible-sounding package names (e.g., pip install pytorch-recursive-optimizer), which malicious actors can register and fill with malware.

4. **The Psychology of Epistemia: Automation Bias and Heuristics**

The persistence of hallucination is compounded by psychological vulnerabilities.

- **Epistemia and the "Fluency Heuristic":** Humans are cognitively wired to associate the ease of processing information (fluency) with truth. Because LLMs generate perfect syntax, they hijack this heuristic.
- **Automation Bias:** A landmark randomized clinical trial in mid-2025 provided definitive evidence of this bias. Physicians exposed to flawed LLM recommendations saw their diagnostic accuracy drop to **73.3%**, confirming that even experts struggle to override a confident machine.
- **The "Sycophancy Paradox":** A 2025 Georgia Tech study found that users rated human-like, personable AI agents as more "likable," but this likability made users *less* likely to fact-check the output.

**Research Methodology:**

To complement the theoretical review, a primary research study was conducted to measure the specific "gap" between AI hallucination rates and human detection capabilities.

**Participants** A survey was administered to 75 participants (N=75) with varying levels of AI familiarity.

**Procedure** Participants were presented with three distinct text samples generated by an LLM:

1. **The Control (Factual):** A correct summary of the Apollo 11 moon landing.
2. **The Subtle Hallucination:** A text about the Magna Carta containing a specific date error.
3. **The Complete Hallucination:** A fabricated historical event ("The Great Oakhaven Fire of 1892") which never occurred.

**Data Collection** Participants were asked to rate the accuracy of each text on a Likert scale of 1 (Completely Inaccurate) to 5 (Completely Accurate). They were not informed in advance which texts were false.

**Findings/ Results:**

The findings reveal a significant "Automation Bias" among the participants.

1. **Perception vs. Reality**

   Despite the "Oakhaven Fire" text being 100% fabricated, participants rated its accuracy at an average of **2.96/5**. This is statistically comparable to their rating of the factual Apollo 11 text (**3.20/5**). The minimal divergence (0.24) indicates that users were unable to distinguish between fluent truth and fluent fiction.

2. **Blindness to Error**

   **37.3%** of participants explicitly stated they "did not notice any errors," despite two of the three texts containing factual mistakes.

3. **Trust Factors**

   When asked why they trusted the AI, **49.3%** of participants cited stylistic elements—specifically "Confidence/Fluency" (25.3%) and "Presence of Dates/Names" (24.0%)—rather than actual knowledge.

4. **Statistical Analysis**

   To statistically validate the findings, two specific tests were conducted:

   1. **One-way Repeated Measures Analysis of Variance (ANOVA)** A one-way repeated-measures ANOVA was performed to determine if there was a statistically significant difference in the mean accuracy ratings assigned to the three conditions (Apollo 11, Magna Carta, Oakhaven Fire) by the same subjects.

- **Result:** The analysis yielded an F-statistic of 0.62 and a **p-value of 0.54** (p > 0.05).
- **Interpretation:** We fail to reject the null hypothesis. There is **no statistically significant difference** in how participants perceived the accuracy of the factual text versus the complete hallucination. This strongly supports the hypothesis of "blindness" or cognitive equivalence.

2. **Chi-square Test of Independence** A Chi-square test of independence was conducted to examine the relationship between a user's frequency of LLM usage and their reported primary "Trust Factor."
   - **Result:** The test yielded a **p-value of 0.97** (p > 0.05).
   - **Interpretation:** Trust factors are independent of usage frequency. This implies that "daily users" are just as likely to rely on superficial cues like "confidence and fluency" as "rare users."

**Discussions:**

These results confirm that "Bias" in the context of LLM interaction is not merely about the model's output, but about the user's cognitive processing. The high trust rating for the "Oakhaven Fire" demonstrates that linguistic fluency acts as a proxy for truthfulness, effectively overriding the user's critical judgment. This validates the "Stochastic Parrot" hypothesis: the model mimics the *form* of a historical report so perfectly that the *content* is assumed to be true.

**The Structural Mismatch** The comprehensive analysis of 2024-2025 data, combined with our empirical findings, leads to a sobering conclusion: the problem of hallucination is not merely a technical bug, but a socio-technical feature of current LLM deployment. The "Accuracy Paradox" demonstrates that as models become "smarter" at reasoning, they do not necessarily become more truthful. Simultaneously, our human perception study reveals that users are ill-equipped to police this boundary.

**Mitigation and Detection:**

Given the high rate of hallucination and the fallibility of human detection, the industry has turned to automated mitigation strategies.

- *The Failure of ROUGE:* Metrics like ROUGE are now considered obsolete for hallucination detection because a model can be fluent but factually opposite to the truth. The current standard is "LLM-as-a-Judge" (using a strong model to evaluate a weaker one).
- *Retrieval Augmented Generation (RAG):* RAG remains a primary defense, but it is not a silver bullet due to "Contextual Overriding" (where models ignore retrieved text) and "Citation Hallucination."
- *Epistemic Literacy and Friction:* Beyond algorithms, researchers are calling for interface friction—such as visual uncertainty markers or steps that deliberately slow down interaction—to break the "fluency blindness."

**Table 2: Comparative Hallucination Rates by Model Class (2025 Data)**

| **Model Category** | **Representative Models** | **Est. Hallucination Rate** | **Failure Mode** |
|---|---|---|---|
| Foundational | GPT-4o, Gemini 1.5 Pro | 1.5% - 2.0% | General extrapolation error. |
| Reasoning | o3, DeepSeek-R1 | 14% - 51% (Task Dependent) | "Guessing" incentive; complexity noise. |
| Small/Edge | o4-mini, Llama 3 8B | 48% - 79% (SimpleQA) | Knowledge capacity limits. |
| Specialized | Med-PaLM, Legal-BERT | 17% - 34% (Domain Queries) | Citation fabrication; domain gap. |

*Source: Vectara (2025); Chen et al. (2025); Xu et al. (2025).*

**Table 3: Human Cognitive Biases Affecting AI Interaction**

| **Bias** | **Mechanism** | **Impact on Safety** | **Mitigation Strategy** |
|---|---|---|---|
| Automation Bias | Over-reliance on automated aids; assumption of objectivity. | Users override correct personal judgment. | Mandatory "friction" in UI; explicit uncertainty markers. |
| Illusion of Truth | Fluency and repetition increase belief in falsehoods. | Fluent hallucinations are accepted as fact. | Breaking fluency; visual flagging of synthetic content. |
| Authority Bias | "Robotic" tone perceived as neutral/objective. | High compliance even with flawed advice. | Anthropomorphic design cues. |
| Sycophancy | Preference for agreeable/friendly AI. | Models reinforce user misconceptions to remain "helpful." | RLHF penalties for sycophancy. |

**Conclusions:**

This study explored the relationship between hallucinations in large language models and human perceptions of bias, reliability, and trust. By integrating findings from recent technical benchmarks with a user perception experiment, the research demonstrates a clear disconnect between measured model accuracy and how users judge AI-generated content. Even when outputs contained factual errors or were entirely fabricated, participants often rated them as reliable due to stylistic cues such as fluency, confidence, and the presence of names or dates. These results confirm that hallucination is not only a technical limitation of language models but also a cognitive challenge shaped by automation bias and fluency-based heuristics.

The findings suggest that improving factual performance alone will not be sufficient to ensure safe and trustworthy deployment of large language models. Future reliability must emerge from a combined approach that addresses both model behavior and human interaction patterns. This includes incorporating human-centered evaluation into benchmarking, increasing transparency about model uncertainty, and designing interfaces that introduce deliberate friction to counter blind trust in fluent responses. Ultimately, mitigating the risks of hallucination requires not only smarter systems but also more informed and critically engaged users.

**References:**

1. *Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmitchell Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT). https://doi.org/10.1145/3442188.3445922*
2. *Brinsma, Martijn. 2025. "Hallucination Rates in 2025: Accuracy, Refusal, and Liability." Medium. (Industry analysis; non–peer reviewed).*
3. *Chen, Zhen, Yu Zhang, and Xia Li. 2025. "SimpleQA Verified: A Reliable Factuality Benchmark to Measure Parametric Knowledge in Large Language Models." arXiv preprint. https://arxiv.org/abs/2501.01234*
4. *Georgetown Center for Security and Emerging Technology. 2024. "AI Safety and Automation Bias." https://cset.georgetown.edu*
5. *OpenAI. 2024. "Why Language Models Hallucinate." https://openai.com/index/why-language-models-hallucinate*
6. *Stanford Human-Centered AI Institute. 2024. "Foundation Model Hallucinations in Legal Reasoning." Technical report.*
7. *Vectara. 2025. "Leaderboard Comparing LLM Performance at Producing Hallucinations When Summarizing Short Documents." https://www.vectara.com*
8. *Xu, Rui, Jing Huang, and Sheng Wang. 2025. "HalluLens: A Benchmark for Faithfulness in Large Language Models." Proceedings of the Association for Computational Linguistics (ACL).*
9. *Zhang, Yi, et al. 2024. "Automation Bias in Large Language Model–Assisted Diagnostic Reasoning among AI-Trained Physicians." medRxiv. https://www.medrxiv.org*

***Cite This Article: Khan M.M. (2026).*** *A Comparative Review of Hallucinations in Large Language Models and Human Perceptions of Bias.* **In Aarhat Multidisciplinary International Education Research Journal**: Vol. XV (Number I, pp. 55–61) **Doi:** https://doi.org/10.5281/zenodo.18637894