INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 1, January 2026 | ISSN: 2456-4184 | IJNRD.ORG

AI Hallucinations in Generative Models: Structural, Data-Driven, and Human Factors Influencing Fabricated Outputs

Munalisa Paul

Abstract

The rapid adoption of generative artificial intelligence has significantly increased reliance on large language models across a wide range of applications. Despite their fluency and apparent confidence, these systems frequently generate hallucinations – outputs that appear plausible but are factually incorrect or fabricated. This paper presents a meta-analytical examination of the underlying causes of AI hallucinations, focusing on architectural, data-driven, and human-interaction factors. It analyses how probabilistic token generation, transformer-based attention mechanisms, evaluation metrics, and context window limitations contribute to hallucinated outputs. The study further examines the role of training data quality, including the amplification of errors through synthetic and imitation datasets, as well as the impact of prompt formulation on inference-time hallucinations. Extending beyond text-based systems, the paper also considers hallucinations in multimodal models, highlighting the increasing difficulty for users to distinguish factual outputs from fabricated ones. The findings demonstrate that AI hallucinations are systemic and predictable outcomes of current training and deployment paradigms rather than isolated errors, underscoring the need for layered mitigation strategies.

Keywords: AI hallucinations, large language models, transformer architectures, training data bias, prompt engineering

Introduction

Generative artificial intelligence (AI) emerged as a concept in the 2010s, but it took a few years to fully come into the limelight and was commercially available to every user by 2022 (Liu, 2024). AI has become a vital part of day-to-day life, with professions such as doctors, engineers, and lawyers utilizing it daily. For instance, research suggests that lawyers use generative AI for important and sensitive tasks such as enhancing their review and drafting of legal documents. According to Statista (2025), generative AI is projected to reach US$59.01bn in 2025. The market size is expected to grow at an annual rate of 37.57%, reaching US$400.00bn by 2031. Hence, generative AI will become the largest global market in the future.

Despite the rapid adoption and impressive capabilities of generative AI models, they are prone to producing fabricated, inaccurate, or illogical information, a phenomenon widely known as hallucination. AI hallucinations are comparable to how humans see figures in the clouds or faces on the moon. In the case of AI, these misinterpretations transpire due to miscellaneous aspects, including overfitting, training data bias/inaccuracy, and high model complexity (Rawte, Sheth, and Das, 2023). Hence, this causes trust issues, especially for professions such as medicine, law, or scientific research. These inaccuracies raise questions for the researchers: How can we reliably detect and reduce AI hallucinations? And what strategies can ensure the safe deployment of generative AI in sensitive domains?

This paper argues that AI hallucinations are not spontaneous or accidental errors but systemic outcomes of data limitations, probabilistic prediction mechanisms, and misaligned training objectives inherent to generative AI models. Although current mitigation strategies reduce hallucinations, they cannot fully eliminate them due to fundamental architectural constraints.

Literature Analysis

AI hallucinations remain a prominent area of discussion amongst researchers and scholars. Some researchers conducted studies, and others identified gaps in how hallucinations are studied across various LLM domains. Pfeiffer et al. (2023) conducted a vital evaluation of translation models, including ChatGPT and M2M neural machine translation models, testing them across 100+ resource settings and translation directions. They found out that hallucinations become very dangerous and harmful in mission-critical fields such as banking, medicine, law, and finance domains, where accuracy is important, and errors and blunders can lead to severe consequences. In medicine, Umapathi et al. (2023) introduced Med-HALT, a benchmark for detecting and reducing hallucinations in medical LLMs. It includes multinational examination data and evaluates both reasoning-based and memory-based hallucinations. In law, Cui et al. (2023) developed ChatLaw, a domain-specific legal LLM supported by a carefully curated legal fine-tuning dataset. To reduce hallucinations in legal information retrieval, they propose a hybrid method that combines vector database search with keyword-based retrieval, improving accuracy over vector search alone. Hence, research on "AI hallucination" has been a substantial topic among scholars from various backgrounds and disciplines for a significant period.

AI Hallucinations are serious matters, and they are not to be considered mere 'errors' from generative AI systems. Differentiation prediction errors, classification errors, and AI hallucinations are a cross-domain phenomenon affecting different types of AI systems. Huang et al. (2024) figured outthat the Errors in Generative AI – specifically in multimodal models and generative models – can occur due to various reasons, and distinguishing among them is very important for an accurate result and evaluation. Prediction errors arise when a model estimates an outcome and produces the wrong output due to insufficient context, statistical uncertainty, or limitations of the training data. These errors occur in probabilistic systems and do not involve fabrication. Classification errors are another type of error that occurs when the model assigns the wrong label to an input due to ambiguous features, domain shifts, or biases in the training distribution; although incorrect, these mistakes remain grounded in the provided input. Hallucinations occur when the model produces clean, confident content that is contradicted or unsupported. Huang et al. (2024) also mentioned that this happens because generative models are fundamentally designed to predict what is statistically probable and also to sound factual. As a result, they may invent shapes, colors, object positions, causal explanations, or entire entities simply because these completions align with patterns learned during training. Zhang (2024) mentioned that this distinction is especially visible in multimodal systems, where a model may fabricate visual attributes (such as incorrectly asserting the shape of cheese, misidentifying colors, or imagining objects that do not exist) despite having direct access to the image. Hence, prediction classifications and prediction mistakes remain within the boundaries of input-driven reasoning. Hallucinations are emerging and proving that the models also have limitations, as the current architectures are limited to the AI modes.

As Liu (2025) said, these hallucinations are not restricted to pure text-based AI; they also occur in multimodal, text-to-image AI systems. As shown in Figure 1, hallucinations can occur when identifying shapes, colors, sizes, and positions. As demonstrated in the provided image, Figure 1, diverse visual scene models produce textual outputs with many different and contradictory aspects regarding colors, shapes, or relationships that do not exist in the actual scene itself. Favero et al. (2024) mentioned that these hallucinatory experiences can arise from the interaction between image analysis and language production. Hallucinations may occur in visual form; however, they are exclusively produced by means of the combination of imaging analysis and written text by way of the language process. Models may invent object properties, spatial relationships, or scene content based on biases or limitations in their visual grounding, resulting in errors that go beyond simple textual mistakes.

Figure 1: The image above compares different shapes, colors, sizes, and positions (Liu, 2025)

Causes of AI Hallucinations

Traditional evaluation frameworks for language models rely heavily on binary metrics such as precision, accuracy, and pass rates. While useful for benchmarking, these metrics inadvertently promote overconfidence by rewarding "best-guess" answers and penalising uncertainty or abstention. Models are forced into a rigid right-or-wrong framework with no space for partial knowledge, probabilistic reasoning, or justified doubt. As a result, responses such as "I don't know" become sub-optimal, even when they would be the most epistemically honest answer. This removes nuance from model assessment and encourages outputs that appear precise and authoritative while masking underlying uncertainty and complexity in reasoning. Consequently, commonly used metrics do not incentivise calibrated or cautious behaviour; instead, they push models toward hallucination rather than uncertainty admission or clarification-seeking (Dhinakaran, 2025).

Transformer architectures themselves also play a significant role in hallucination behaviour. Research on vision transformers demonstrates that neuron activations can develop detectable semantic structure across multiple layers, enabling the identification and manipulation of interpretable concepts. Interpretability is measured through the semantic similarity of labels associated with the top images that maximally activate each concept, while steerability is evaluated through causal interventions—specifically, injecting a concept into the residual stream of neutral images to induce a corresponding class label (Dhinakaran, 2025). Building on this, Liu (2023) showed that a large proportion of noise-derived concepts extracted using sparse autoencoders (SAEs) are highly interpretable, with many, particularly in early and mid-layers, also demonstrating steerability.

To test the robustness and organisational structure of these concepts, multiple SAE models were trained on identical layer activations using different random seeds and configurations. Across runs, a stable core of concepts repeatedly emerged in the same layers, suggesting non-random internal organisation. When comparing pairs of SAEs, a three-phase overlap pattern becomes evident. In the initial layers, approximately 35% of concepts overlap, reflecting strong constraints imposed by input data and limited architectural bias. This overlap drops to around 14% in intermediate layers, where the model explores alternative semantic interpretations. In deeper layers, overlap increases again to roughly 25%, indicating convergence toward stable, high-level representations. This progression suggests that transformers gradually refine noisy activations into task-relevant features that remain consistent across training variations (Suresh et al., 2023).

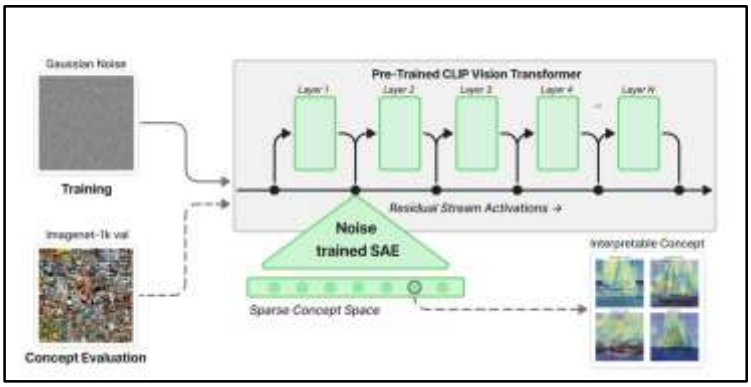

Figure 2 (Suresh et al., 2023)

Figure 2 illustrates this process by showing how researchers uncover hidden "visual concepts" learned by a CLIP Vision Transformer using a sparse autoencoder. The SAE is trained exclusively on random noise rather than real images, allowing it to identify generalised patterns rather than memorising specific visual details. Once trained, real images are passed through the model, and individual SAE neurons activate in response to particular features. This reveals the types of visual concepts encoded internally and provides a window into the model's representational structure, helping to make its internal reasoning processes more interpretable.

At a functional level, transformers generate outputs by attending to previous tokens and predicting the most probable next token based on learned statistical patterns. A hallucination-prone transformer does not differ architecturally from a well-performing one; instead, the difference lies in behavioural conditions. When prompts are ambiguous, incomplete, or fall outside the model's training distribution, next-token prediction becomes weakly constrained. In such cases, the model continues producing fluent sequences that sound coherent but lack factual grounding. This occurs because self-attention mechanisms optimise for linguistic plausibility rather than truth verification. Consequently, hallucinations arise not because the model "knows" incorrect facts, but because it extends statistical regularities beyond their valid domain, fabricating details, citations, or reasoning chains that align with learned patterns rather than external reality.

These issues are further compounded by structural limitations inherent to transformer-based systems. Attention weights tend to decay or become diffuse over long contexts, causing earlier constraints to lose influence as generation progresses. In extended or open-ended tasks, the model gradually fills informational gaps with statistically likely but unverified content. The same pattern-matching objective also drives overgeneralisation: models learn correlations rather than grounded rules and therefore extend patterns beyond appropriate contexts. Since the training objective rewards plausibility under the data distribution rather than explicit truthfulness or uncertainty calibration, the model lacks an internal mechanism to reliably distinguish between "plausible-sounding" and "correct." As a result, hallucinations emerge as a predictable byproduct of transformer architectures and objectives, persisting even when outputs appear confident, coherent, and well-structured.

Training data quality further reinforces this phenomenon. Large language models are only as reliable as the data they are trained on, and internet-scale datasets inevitably contain biases, misinformation, outdated facts, and internal inconsistencies (Li et al., 2023). Because LLMs learn patterns rather than verified truths, these imperfections become embedded within their internal representations and can later surface as confident but incorrect outputs. Hallucinations are especially likely when the training data sparsely covers a topic: instead of expressing uncertainty, the model predicts the most statistically probable continuation, effectively inventing details to maintain fluency. An emerging concern is the increasing use of synthetic or AI-generated text in training pipelines, which risks amplifying existing errors through recursive feedback loops. Over time, small inaccuracies can be repeated, reinforced, and normalised within newer models.

Foundational work by Brown et al. (2020) demonstrated that GPT-style models are trained solely through next-token prediction and can generate highly coherent and contextually appropriate text without any explicit

mechanism for fact-checking. Because likelihood and truth are not equivalent, models can produce statements that sound correct but are factually false. Subsequent work on reinforcement learning from human feedback (RLHF) showed that while instruction-following and politeness improve, hallucinations persist. RLHF often incentivises responses that humans prefer, confident, fluent, and decisive, rather than responses that are strictly accurate (Ouyang et al., 2022). Benchmarks and human evaluators similarly tend to reward certainty over epistemic humility, reinforcing confident guessing. Collectively, this research indicates that hallucinations are not accidental defects but predictable outcomes of training regimes that prioritise human preference and fluency over factual correctness (F Christiano et al., 2017)

Inference-time factors further contribute to hallucination. Laubheimer (2025) investigated how unclear or underspecified prompts can trigger the generation of plausible but unsupported content. Prompt formulation plays a critical role: even minor ambiguity can lead the model to infer intent incorrectly and fabricate details to fill perceived gaps. Empirical observations suggest that prompt-related errors contribute significantly to hallucination frequency, with some studies indicating incorrect outputs occurring in roughly one out of six interactions under certain conditions. Figure 1.1 illustrates this effect through a case study in which ChatGPT incorrectly identified an album cover from an image when given a minimal prompt. Although the model initially acknowledged uncertainty, a step toward hallucination mitigation, it ultimately produced a confident but incorrect identification (Laubheimer, 2025).

Figure 3

Context window limitations represent another structural contributor. A model's context window defines the maximum number of tokens it can process at once and functions as a form of short-term memory (Synced, 2024). When inputs exceed this limit, earlier information may be truncated, leading to loss of critical constraints. Even within large context windows, models often exhibit the "murky middle" problem, disproportionately focusing on the beginning and end of long texts while neglecting details in the middle. This degrades performance on tasks such as long-document summarisation and increases the likelihood of fabricated connections or omissions (Data Science Dojo, 2024; Rinko, 2024).

Overall, AI hallucinations stem from systemic and structural factors rather than isolated technical failures. They emerge from the interaction between probabilistic token generation, training objectives that reward plausibility over truth, imperfect and biased data, prompt ambiguity, and memory limitations. Importantly, hallucinations are not solely a property of the model itself; they are equally shaped by how users interact with the system and the quality of data it has been exposed to. Reducing hallucinations, therefore, requires a multi-layered approach –

improvements in model architecture and training objectives, stronger data curation practices, and more careful prompt design – rather than treating hallucination as a simple bug to be patched.

Conclusion

Since 2010, generative AI has gradually emerged as a significant technological concept, gaining widespread visibility and commercial accessibility by 2022. Since then, the generative AI market has expanded rapidly, with many users increasingly relying on AI systems in their daily activities. Despite this growth, generative AI systems – particularly conversational agents – continue to produce fabricated outputs that may sound convincing but are factually incorrect. Such hallucinations can mislead users, especially when the generated responses appear confident and authoritative.

This paper identified several key factors that contribute to AI hallucinations, including the quality and structure of training data, tokenization and probabilistic token generation, limitations within transformer architectures, and human-related factors such as poorly formulated prompts. Addressing these issues has the potential to significantly reduce hallucinated content. Improvements in prompt formulation, data curation, and tokenization processes, alongside architectural refinements, can help ensure the safer deployment of generative AI systems, particularly in sensitive and high-stakes domains. Of particular concern is the growing reliance on artificial or imitation data, content generated by other AI models for training, risking the reinforcement of earlier errors and creating feedback loops in which inaccuracies are repeatedly amplified. Ultimately, the reliability of large language models is closely tied to the quality of their training data; when that data is flawed or inconsistent, the model's outputs inevitably reflect those same limitations.

Because AI hallucination is a systemic rather than isolated error, it arises from multiple interacting layers within the model and its usage context. Accordingly, this paper conducted a meta-analysis examining the nature of AI hallucinations and the cumulative factors that lead to their occurrence, spanning both architectural and multimodal dimensions. Hallucinations are not confined to text-based outputs alone but also occur in image and multimodal generation, making it increasingly difficult for users to distinguish between factual and fabricated content. This presents a growing challenge as AI systems become more deeply embedded in everyday decision-making. The findings of this paper highlight an important gap that can be addressed by correcting issues at the model-layer level, as well as by improving human–AI interaction through clearer and more precise prompting. By tackling these factors collectively, the frequency and impact of AI hallucinations can be meaningfully reduced.

Bibliography

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C. and Hesse, C. (2020). *Language Models are Few-Shot Learners*. [online] Available at: https://arxiv.org/pdf/2005.14165.

Cui, J., Li, Z., Yang, Y., Chen, B. and Liu, Y. (2023). ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. *arXiv (Cornell University)*. doi:https://doi.org/10.48550/arxiv.2306.16092.

Data Science Dojo (2024). *LLM Context Window Paradox: 5 Ways to Solve the Problem*. [online] Data Science Dojo. Available at: https://datasciencedojo.com/blog/the-llm-context-window-paradox/.

Dhinakaran, A. (2025). *Testing Binary vs Score Evals on the Latest Models*. [online] Arize AI. Available at: https://arize.com/blog/testing-binary-vs-score-llm-evals-on-the-latest-models/ [Accessed 18 Dec. 2025].

F Christiano, P., Leike, J., Brown, T.B., Martic, M., Legg, S. and Amodei, D. (2017). *Deep Reinforcement Learning from Human Preferences*. [online] Available at:

https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Favero, A., Luca Zancato, Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A. and Stefano Soatto (2024). *Multi-modal hallucination control by visual information grounding*. [online] Amazon Science. Available at: https://www.amazon.science/publications/multi-modal-hallucination-control-by-visual-information-grounding [Accessed 18 Dec. 2025].

Huang, W., Liu, H., Guo, M. and Gong, N. (2024). *Visual Hallucinations of Multi-modal Large Language Models*. [online] pp.9614–9631. Available at: https://aclanthology.org/2024.findings-acl.573.pdf [Accessed 18 Dec. 2025].

Laubheimer, P. (2025). *AI Hallucinations: What Designers Need to Know*. [online] Nielsen Norman Group. Available at: https://www.nngroup.com/articles/ai-hallucinations/.

Li, T., Cheng, L., Hosseini, M., Johnson, M. and Steedman, M. (2023). *Sources of Hallucination by Large Language Models on Inference Tasks*. [online] Available at: https://arxiv.org/pdf/2305.14552.

Liu, B. (2024). *How Is Generative Artificial Intelligence Changing the Legal profession? - Economics Observatory*. [online] Economics Observatory. Available at: https://www.economicsobservatory.com/how-is-generative-artificial-intelligence-changing-the-legal-profession.

Liu, F. (2025). *Digital Repository at the University of Maryland (DRUM)*. [online] Umd.edu. Available at: https://drum.lib.umd.edu/items/d40cfa16-f261-4e14-950a-9875b08abf49 [Accessed 18 Dec. 2025].

Liu, X. (2023). *From Noise to Narrative: Tracing the Origins of Hallucinations in Transformers*. [online] Arxiv.org. Available at: https://arxiv.org/html/2509.06938v1 [Accessed 18 Dec. 2025].

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. and Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv:2203.02155 [cs]*. [online] Available at: https://arxiv.org/abs/2203.02155.

Pfeiffer, J., Piccinno, F., Nicosia, M., Wang, X., Reid, M. and Ruder, S. (2023). mmT5: Modular Multilingual Pre-Training Solves Source Language Hallucinations. *arXiv (Cornell University)*. doi:https://doi.org/10.48550/arxiv.2305.14224.

Rawte, V., Sheth, A.P. and Das, A. (2023). A Survey of Hallucination in Large Foundation Models. *arXiv (Cornell University)*. doi:https://doi.org/10.48550/arxiv.2309.05922.

Rinko, S. (2024). *6 Best Large Language Models (LLMs) in 2024*. [online] eWEEK. Available at: https://www.eweek.com/artificial-intelligence/best-large-language-models/.

Statista (2025). *Generative AI - Worldwide | Statista Market Forecast*. [online] Statista. Available at: https://www.statista.com/outlook/tmo/artificial-intelligence/generative-ai/worldwide?srsltid=AfmBOoon33y5vx-_397uqt3O6445xfMxsK0tHlFF14jCbSnDLS3RnVUp [Accessed 15 Dec. 2025].

Suresh, P., Stanley, J., Joseph, S., Scimeca, L. and Bzdok, D. (2023). *From Noise to Narrative: Tracing the Origins of Hallucinations in Transformers*. [online] Arxiv.org. Available at: https://arxiv.org/html/2509.06938v1 [Accessed 18 Dec. 2025].

Synced (2024). *Microsoft's LongRoPE Breaks the Limit of Context Window of LLMs, Extents it to 2 Million Tokens | Synced*. [online] Synced | AI Technology & Industry Review. Available at: https://syncedreview.com/2024/02/25/microsofts-longrope-breaks-the-limit-of-context-window-of-llms-

extents-it-to-2-million-tokens/ [Accessed 18 Dec. 2025].

Umapathy, V.R., B, S.R., Raj, R.D.S., Yadav, S., Munavarah, S.A., Anandapandian, P.A., Mary, A.V., Padmavathy, K., R, A., Umapathy, V.R., B, S.R., Rajkumar, D.S.R., Yadav, S., Munavarah, S.A., Iv, P.A.A., Mary, V., Padmavathy, D.K. and R, A. (2023). Perspective of Artificial Intelligence in Disease Diagnosis: A Review of Current and Future Endeavours in the Medical Field. *Cureus*, [online] 15(9). doi:https://doi.org/10.7759/cureus.45684.

Zhang, T. (2024). *Multi-modal Attribute Prompting for Vision-Language Models*. [online] Arxiv.org. Available at: https://arxiv.org/html/2403.00219v3 [Accessed 18 Dec. 2025].