Journal of Intelligent Computing

DLINE JOURNALS

Print ISSN: 0976-9005
Online ISSN: 0976-9013

JIC 2026: 17 (4)
https://doi.org/10.6025/jic/2026/17/4/179-199

Multidimensional Hallucination Risk Across AI Models: Frequency, Confidence, Accuracy Mismatch, Detection, and Trust Impact

Pit Pichappan
Digital Information Research Labs
Chennai. Tamil Nadu India
pichappan@dirf.org

ABSTRACT

Large language models (LLMs) frequently generate plausible but factually inaccurate outputs, commonly known as hallucinations. Moving beyond treating hallucination as a simple binary factuality problem, this study evaluates it as a multidimensional risk encompassing generation frequency, confidence accuracy mismatch, detectability, and downstream trust impact. Analyzing a comprehensive dataset of 10,000 interactions across six frontier AI models and platforms, we assessed hallucination prevalence, epistemic calibration, and human consequences. Results reveal distinct variations in hallucination rates, with Llama and Mistral exhibiting the highest frequencies. Crucially, we identified a pronounced confidence accuracy mismatch across systems; notably, Mistral presented the most concerning risk configuration by combining high hallucination rates, maximum AI confidence, and the lowest detection rates. Conversely, GPT-5 demonstrated the most favorable factual accuracy profile. Cross-tool analysis demonstrated that hallucination occurrence, rather than specific tool identity, serves as the primary driver of severe trust erosion. Specifically, the mere presence of an error caused a massive trust penalty that subsequent detection alone could not resolve. These findings proved that AI hallucination is a systemic, cross-platform reliability challenge. Mitigating its impact requires integrated strategies addressing confidence calibration, proactive detection mechanisms, and user interface designs that foster appropriate skepticism, rather than relying solely on reducing raw factual error rates.

Keywords: Large Language Models, AI Hallucination, Multidimensional Risk, Confidence Accuracy Mismatch, Trust Erosion, Error Detection, Model Reliability, Epistemic Calibration, Risk Configuration, Cross Tool Consistency

Received: 18 April 2026, Revised: 15 July 2026, Accepted: 3 August 2026

Copyright: DLINE

1. Introduction and Background

Artificial intelligence (AI) represents a paradigm shift in delivering data driven, innovative knowledge, but inaccurate or unverifiable information can substantially influence how users interpret information and make decisions. Consequently, understanding hallucination related risks is essential for evaluating the trustworthiness and ethical implications of AI systems used in decision making [1].

In large language models (LLMs), hallucinations generally refer to instances in which generated content appears coherent, plausible, and linguistically well formed but contains factual inaccuracies or unverifiable claims [2, 3,4]. Although hallucinations are commonly discussed as limitations that can be reduced through improved data, training, prompting, or model design, another perspective views them as closely connected to fundamental properties of probabilistic language generation. From this perspective, training LLMs for predictive accuracy can still produce hallucinated outputs irrespective of model architecture or data quality because language generation is probabilistic and because LLMs face inherent limitations in learning all computable functions [5, 4]. These limitations are further linked to constraints identified in computational learning theory and to the absence of complete mappings between language, knowledge, and factual reality [6].

Modern LLMs can therefore generate unexpected, fabricated, or unsupported details that affect how humans interpret, evaluate, and rely on AI-generated information. This issue becomes particularly consequential in high-stakes settings, including medical triage, legal documentation, and policy recommendations, where inaccurate information may lead to substantial downstream consequences. Although quantitative evaluations of hallucination commonly focus on factual inconsistencies in model outputs, these errors' effects on human users and their subsequent decisions have received comparatively less systematic attention [7]. A comprehensive assessment of hallucination risk should consequently extend beyond whether an output is factually correct and consider how confidently information is presented, whether users or systems can detect errors, and how hallucinations ultimately affect trust in AI-generated information.

Research on LLM hallucinations has expanded rapidly and now encompasses a broad range of perspectives concerning their causes, manifestations, detection, and mitigation. Several surveys have provided comprehensive accounts of hallucination mitigation across different stages of the LLM pipeline, extending from pre-generation interventions to post-generation verification and correction [8, 9, 10, 11, 12]. Existing literature has often adopted task-oriented taxonomies that organise mitigation approaches by downstream applications, such as summarisation, question answering, and code generation [13], or according to the level at which an intervention is applied, including pre-training, fine-tuning, and decoding [14]. These approaches have substantially improved understanding of where hallucinations occur and which applications are particularly vulnerable; however, they offer less insight into the underlying methodological principles behind hallucination mitigation.

To address this limitation, method oriented taxonomies have increasingly classified mitigation strategies by core techniques, including data augmentation, retrieval integration, prompt engineering, and external fact-checking [15, 16]. Within this broader methodological landscape, several studies have concentrated on specific mitigation families. Retrieval augmented generation has been investigated as a means of grounding model responses in retrieved information [17, 18], while other research has examined ambiguity and uncertainty estimation [19], knowledge graph based approaches [20], and multi-agent reinforcement learning [21]. Such studies provide valuable and detailed insights into particular mitigation mechanisms, although their focus on individual classes of techniques can make it difficult to develop an integrated understanding of hallucination risk across generation, confidence, detection, and user trust [22].

Prompting is another important determinant of hallucination behaviour. Poorly specified or insufficient prompts can increase hallucination risk because they omit information needed to constrain the model's response or leave the intended task ambiguous [23]. Under such conditions, the model may rely more heavily on probabilistic pattern matching, favouring broad generalisation rather than closely aligning its response with the user's specific intent [24, 25, 26]. Adversarial and out of distribution prompts can further increase

vulnerability by presenting inputs that differ substantially from patterns encountered during model development [27, 28].

Hallucination risk may also increase in long context settings. As the amount of information presented to a model grows, inaccuracies can emerge from summarisation or retrieval failures during extended-sequence attention, potentially causing the model to lose grounding or relevant contextual information over time [29, 30]. These challenges matter most when models must integrate information across long documents or multiple sources. At the same time, resource constrained environments may lack access to computationally intensive mitigation approaches such as fine tuning. In such settings, zero shot detection techniques, including self-consistency probes [31] and hesitation analysis [32], provide potential mechanisms for identifying hallucinations without substantial additional computational overhead.

The implications of hallucination extend beyond technical model performance into academic and scholarly contexts. Hallucinations can systematically compromise academic writing quality, distort assessment processes, and undermine epistemic trust in scholarly outputs [33]. Importantly, variation in hallucination rates across models and disciplines indicates that hallucination behaviour is context dependent rather than uniform. Several interacting factors may contribute to this variation, including probabilistic text generation, limitations in training data, insufficient contextual understanding, and the lack of robust mechanisms to verify generated information. Recent work has also broadened hallucination research from text only LLMs to multimodal AI systems. Liu [34] presented a comprehensive evaluation of hallucination phenomena across LLMs and multimodal systems and proposed a structured taxonomy incorporating factuality based, faithfulness based, logical based, and emergent hybrid forms. The framework further recognises multimodal specific risks, including cross modal inconsistencies, visual overinterpretation, and modality dominance effects. Beyond categorising hallucination types, this work traces hallucinations across the model lifecycle, identifying data-level origins such as knowledge gaps, misinformation, and annotation noise; training-induced mechanisms such as distributional mimicry, reward bias, and alignment forgetting; and inference-time vulnerabilities including confidence miscalibration, decoding failures, and prompt-induced errors. [35]

The above discussions indicate that hallucination should not be treated solely as a binary factuality problem. It is a multidimensional risk involving how often inaccurate information is generated, how well model confidence matches actual accuracy, users' or automated systems' ability to detect errors, and the consequences of hallucinations for trust and decision making. This broader perspective matters because a highly fluent, confident but incorrect response may be more consequential than an obviously uncertain error. Accordingly, an integrated examination of hallucination frequency, confidence accuracy mismatch, detection behaviour, and trust impact can provide a more comprehensive assessment of AI reliability than isolated factuality measures alone.

The present study is therefore positioned within this broader research landscape by treating hallucination as a multidimensional phenomenon rather than a single output level error. By considering the relationship between hallucination occurrence, confidence, detection, and trust, the analysis seeks to connect technical characteristics of AI-generated responses with their potential human and decision making consequences. This approach aligns with the growing recognition that improving AI reliability requires not only reducing hallucinations but also understanding how confidently they are expressed, how effectively they can be identified, and how they influence users' willingness to trust and rely on AI systems.

2. Problem Statement

Large language models (LLMs) have rapidly become integral to knowledge work, decision support, and everyday information seeking. Despite their fluency and apparent competence, these systems frequently generate outputs that are coherent, plausible, and linguistically well formed yet factually inaccurate or unverifiable, a phenomenon widely termed hallucination. Hallucinations are not merely isolated technical failures; they constitute a multidimensional risk that affects how users interpret information, calibrate their confidence, detect errors, and ultimately decide whether to trust and act upon AI-generated content.

Existing research has predominantly treated hallucination as a binary factuality problem or has focused on mitigation techniques at specific stages of the model pipeline (pre-training, fine tuning, decoding, or retrieval augmentation). While valuable, such approaches provide an incomplete picture of real world risk. A highly fluent and confidently expressed incorrect response may pose greater practical danger than an obviously uncertain or easily detectable error. Consequently, a comprehensive assessment of hallucination risk must jointly examine four interrelated dimensions: (1) the frequency with which hallucinations occur, (2) the degree of mismatch between model confidence and factual accuracy, (3) the detectability of erroneous outputs by users or automated systems, and (4) the downstream impact of hallucinations on user trust and decision-making.

This study addresses the lack of an integrated, multi-dimensional empirical comparison of contemporary AI models that simultaneously quantifies hallucination prevalence, confidence accuracy calibration, detection performance, harm potential, and trust erosion. Without such an integrated view, practitioners, policymakers, and end users lack the evidence needed to evaluate relative model reliability, to prioritise safety interventions, or to design appropriate human–AI interaction protocols for high-stakes domains such as medicine, law, education, and public policy.

2.1 Research Issues and Questions

Building on the problem formulation above, the present study is guided by the following research issues:

1. Frequency heterogeneity: Do contemporary frontier and open-weight LLMs differ meaningfully in the rate at which they produce hallucinated outputs, and what is the practical magnitude of any observed differences?

2. Confidence–accuracy mismatch: To what extent do models express high confidence even when factual accuracy is relatively low, and does the severity of this mismatch vary systematically across models?

3. Detectability of hallucinations: Once a hallucination occurs, how reliably is it detected, and do detection rates and detection times differ across models?

4. Joint risk configuration: Which models exhibit particularly concerning combinations of high hallucination frequency, high confidence, and low detectability?

5. Trust impact: How strongly is hallucination occurrence associated with reductions in user trust, and does detection of the hallucination substantially mitigate the trust penalty?

6. Cross-tool consistency: Are the patterns observed at the model level also visible when responses are aggregated by AI tool or platform, or does tool identity itself explain substantial variance in hallucination-related outcomes?

Collectively, these issues motivate an empirical design that moves beyond single-metric model rankings and instead characterises each system by its multidimensional risk profile.

3. Methods

3.1 Dataset and Scope

The analysis utilised the complete AI Hallucination Experience Dataset comprising 10,000 interaction records. Six models were represented in the primary model level analysis: GPT-5 (n = 2,034), Gemini (n = 1,838), GPT-4o (n = 1,800), Claude (n = 1,691), Mistral (n = 1,475), and Llama (n = 1,162). A complementary tool-level analysis aggregated interactions into six platform categories (ChatGPT, Claude, Gemini, Local LLM, Microsoft Copilot, and Perplexity) to examine whether observed patterns generalised beyond individual model identifiers.

3.2 Operational Definitions and Measures

The following measures were derived from the dataset:

1. Hallucination rate: proportion of interactions classified as containing a hallucination (hallucinated interactions ÷ total interactions for the model or tool).

2. Factual accuracy: mean rated factual correctness of the model's output (scale documented in the source dataset).

3. AI confidence: mean self-reported or observed confidence associated with the generated response.

4. Detection rate: proportion of hallucinated interactions that were subsequently identified as erroneous (computed only among hallucinated cases).

5. Detection time: mean time (seconds) required to detect a hallucination, summarised only for cases in which detection occurred.

6. Verification rate: proportion of all interactions in which the user or system performed verification of the output.

7. Harm severity: ordinal classification of potential harm among hallucinated interactions (Low = 1, Moderate = 2, High = 3, Critical = 4); high/critical harm proportions were also reported.

8. Trust reduction: difference between user trust rating before and after the interaction (User Trust Before " User Trust After). We computed separate summaries for all interactions and for hallucinated interactions only.

3.3 Analytical Approach

We first computed descriptive statistics (rates, means, and proportions) for each model and tool. Categorical outcomes (hallucination occurrence, detection among hallucinations, verification, and harm category) were compared using chi-square tests of independence, with Cramer's V reported as a measure of association strength. We examined continuous outcomes (factual accuracy, AI confidence, detection time, and trust reduction) using classical ANOVA, Welch ANOVA (robust to heteroscedasticity), and the non-parametric Kruskal Wallis test. Effect sizes were quantified using omega squared (ω^2). Where omnibus tests indicated significant differences, pairwise post hoc comparisons were conducted with Holm adjustment for multiple testing, and Hedges' g was reported as a standardised effect size.

In addition to single metric comparisons, we constructed an exploratory model × hallucination risk profile index. The index is a rank based composite that jointly considers high hallucination rate, high AI confidence, and low detection rate. It is intended solely as a comparative screening indicator and is not presented as a validated absolute safety score.

Finally, a conceptual pathway model—AI Tool → Hallucination → Detection → Harm →Trust was examined descriptively and with formal tests to determine whether tool identity or hallucination occurrence itself carried greater explanatory power for downstream trust outcomes.

4. Model-level results

4.1 Hallucination Analysis

Hallucination frequency. Llama showed the highest hallucination rate (34.51%), closely followed by Mistral (34.37%). Gemini was intermediate (30.79%), GPT-4o was 29.17%, GPT-5 was 28.17%, and Claude had the lowest rate (27.62%).

Accuracy and confidence. Mean factual accuracy was highest for GPT-5 (7.15), followed by Claude (7.13), GPT-4o (7.09), Gemini (7.01), Llama (6.89), and Mistral (6.84). Mean AI confidence was high for every model, ranging from 7.89 for GPT-5 to 8.00 for Mistral.

Detection and verification. Among hallucinated interactions, Llama had the highest detection rate (59.35%), followed by GPT-4o (58.86%), GPT-5 (57.42%), Claude (56.10%), Gemini (55.48%), and Mistral (53.65%). Mean detection time ranged from 36.92 seconds for Mistral to 41.65 seconds for Llama. Verification rates across all interactions ranged from 38.71% to 40.27%.

Harm and trust. Mean ordinal harm among hallucinations ranged from 1.75 to 1.82. The high/critical harm proportion was highest for GPT-4o (22.29%) and lowest for Mistral (16.96%). Across all interactions, mean trust reduction was highest for Llama (0.75) and Mistral (0.70) and lowest for Claude (0.60). Among hallucinated interactions, trust reduction ranged from 1.76 for GPT-4o to 1.90 for Llama.

AI Model	N	Hallucination rate %	Factual accuracy mean	AI confidence mean	Detection rate %	Detection time mean sec (detected)	Verification rate %	Trust reduction mean
Llama	1162.00	34.51	6.89	8.00	59.35	41.65	38.90	0.75
Mistral	1475.00	34.37	6.84	8.00	53.65	36.92	38.71	0.70
Gemini	1838.00	30.79	7.01	7.94	55.48	38.61	38.90	0.66
GPT-4o	1800.00	29.17	7.09	7.89	58.86	40.54	40.17	0.61
GPT-5	2034.00	28.17	7.15	7.89	57.42	38.09	38.50	0.63
Claude	1691.00	27.62	7.13	7.94	56.10	38.06	40.27	0.60

Table 1. Model-level comparison

4.2 Statistical testing

Categorical outcomes. Hallucination occurrence differed significantly by model, $x^2(5)=32.68$, $p<0.001$, Cramer's $V=0.057$, indicating a small association. Verification did not differ, $x^2(5)=2.19$, $p=0.822$, $V=0.015$. When evaluated only among hallucinated interactions, detection did not differ significantly, $x^2(5)=4.60$, $p=0.467$, $V=0.039$. Harm category among hallucinations also did not differ significantly, $x^2(15)=17.78$, $p=0.275$, $V=0.044$.

Continuous outcomes. Factual accuracy showed a significant model effect under classical ANOVA, Welch ANOVA and Kruskal Wallis testing. AI confidence, detection time and trust reduction restricted to hallucinated cases did not show significant omnibus differences. Trust reduction across all interactions showed a statistically significant model effect, but its effect size was very small.

4.3 Model × hallucination risk profile

Risk interpretation. Mistral produced the clearest high-risk configuration: a 34.37% hallucination rate, the highest mean confidence (8.00), and the lowest hallucination detection rate (53.65%). Llama also had a very high hallucination rate (34.51%) and high confidence (8.00), but its detection rate was higher at 59.35%. Gemini showed a second concerning combination of relatively high hallucination frequency (30.79%), high confidence (7.94), and comparatively low detection (55.48%). GPT-5 showed the most favourable accuracy profile, with the lowest hallucination rate (28.17%) and highest factual accuracy (7.15).

Outcome	ANOVA F	ANOVA p	Welch F	Welch p	Kruskal-Wallis H	Kruskal-Wallis p	Omega squared
Factual accuracy	7.6136	0.0000	7.3616	0.0000	30.4301	0.0000	0.0033
AI confidence	1.8537	0.0989	1.8641	0.0972	6.9950	0.2210	0.0004
Detection time (detected hallucinations)	1.5814	0.1621	1.5311	0.1777	6.4532	0.2646	0.0017
Trust reduction	5.2909	0.0001	5.0280	0.0001	30.8418	0.0000	0.0021
Trust reduction (hallucinations)	1.4694	0.1964	1.5043	0.1854	8.8568	0.1149	0.0008

Table 2. Omnibus model-comparison tests

Outcome	Model 1	Model 2	Holm-adjusted p	Hedges g
Factual accuracy	Mistral	GPT-5	0.0000	-0.1679
Factual accuracy	Claude	Mistral	0.0002	0.1565
Factual accuracy	Llama	GPT-5	0.0022	-0.1409
Factual accuracy	Mistral	GPT-4o	0.0025	-0.1313
Factual accuracy	Llama	Claude	0.0086	-0.1296

Table 3. Significant post-hoc comparisons after Holm correction

AI Model	Halluci-nation rate %	AI confidence mean	Detection rate %	Risk profile index
Mistral	34.373	7.997	53.649	0.889
Gemini	30.794	7.944	55.477	0.667
Llama	34.509	7.995	59.352	0.611
Claude	27.617	7.937	56.103	0.389
GPT-4o	29.167	7.891	58.857	0.333
GPT-5	28.171	7.889	57.417	0.278

Table 4. Exploratory model × hallucination risk profile

The risk-profile index is a rank-based exploratory screening indicator constructed from high hallucination, high confidence and low detection. It is not a validated absolute safety score.

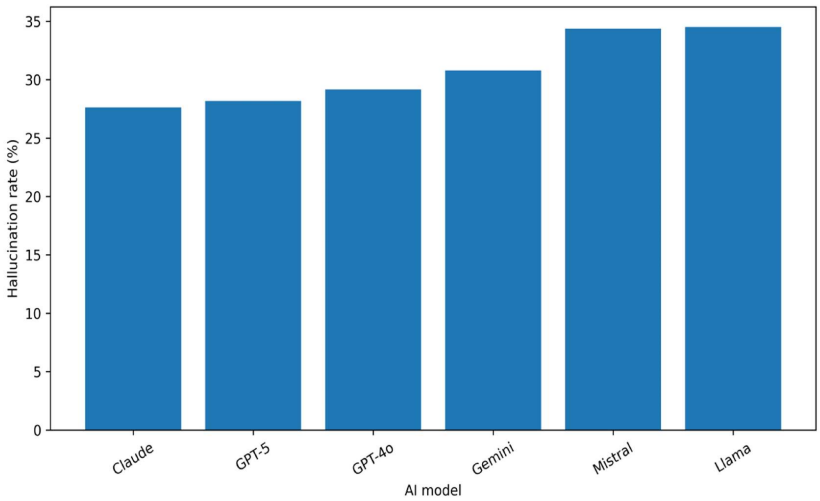

Figure 1. Hallucination rate by AI model

Figure 1 illustrates the variation in hallucination rates across the six evaluated AI models. Llama recorded the highest hallucination rate at 34.51%, followed very closely by Mistral at 34.37%. Gemini showed an intermediate hallucination rate of 30.79%, while GPT-4o recorded 29.17%. GPT-5 had a lower hallucination rate of 28.17%, and Claude exhibited the lowest rate at 27.62%.

The figure indicates that hallucination occurrence was not uniform across models. The difference between the highest and lowest rates was about 6.9 percentage points, suggesting meaningful variation in how often models generated hallucinated responses. However, the statistical analysis showed that this model-level difference, although statistically significant, had a small association effect ($x^2(5)=32.68$, $p<0.001$, Cramer's $V=0.057$).

The figure therefore shows that hallucination frequency alone does not fully assess model reliability. In particular, Llama's highest hallucination rate should be interpreted together with its relatively high hallucination detection rate, whereas Mistral's similarly high hallucination frequency is accompanied by poorer detection performance.

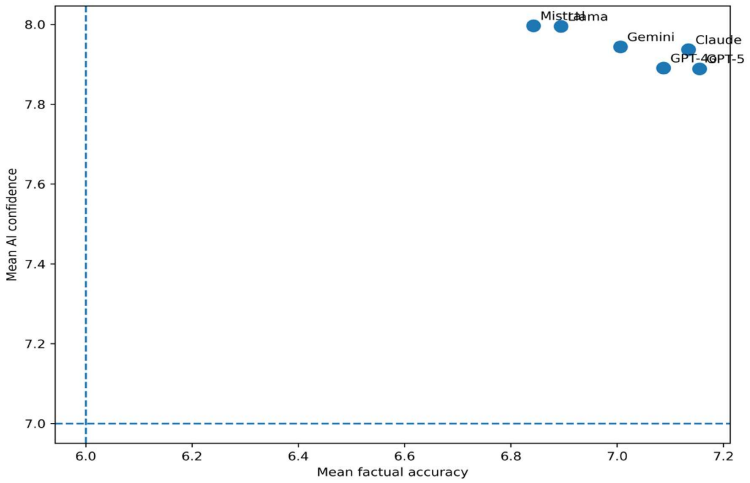

Figure 2. Factual accuracy versus AI confidence by model

Figure 2 compares the mean factual accuracy and AI confidence scores for the six models, providing an important view of the relationship between the correctness of generated information and the confidence associated with those outputs. GPT-5 achieved the highest mean factual accuracy (7.15), followed by Claude (7.13), GPT-4o (7.09), Gemini (7.01), Llama (6.89), and Mistral (6.84).

In contrast, AI confidence remained consistently high across all models, ranging from 7.89 for GPT-5 and GPT-4o to approximately 8.00 for Mistral and Llama. This produces an important discrepancy: models with comparatively lower factual accuracy did not necessarily express lower confidence. Mistral is particularly notable because it combined the lowest factual accuracy (6.84) with very high confidence (8.00).

Statistical testing confirmed that factual accuracy differed significantly between models across ANOVA, Welch ANOVA, and Kruskal Wallis procedures (all $p<0.001$), although the effect was small ($\omega^2=0.0033$). AI confidence, however, did not differ significantly among models (ANOVA $p=0.0989$; Welch $p=0.0972$; Kruskal–Wallis $p=0.2210$).

Thus, Figure 2 provides evidence of a potential confidence–accuracy mismatch. High confidence should not automatically be interpreted as evidence of high factual reliability. This is especially important for hallucination risk assessment because confidently expressed incorrect information may be more difficult for users to recognize as erroneous.

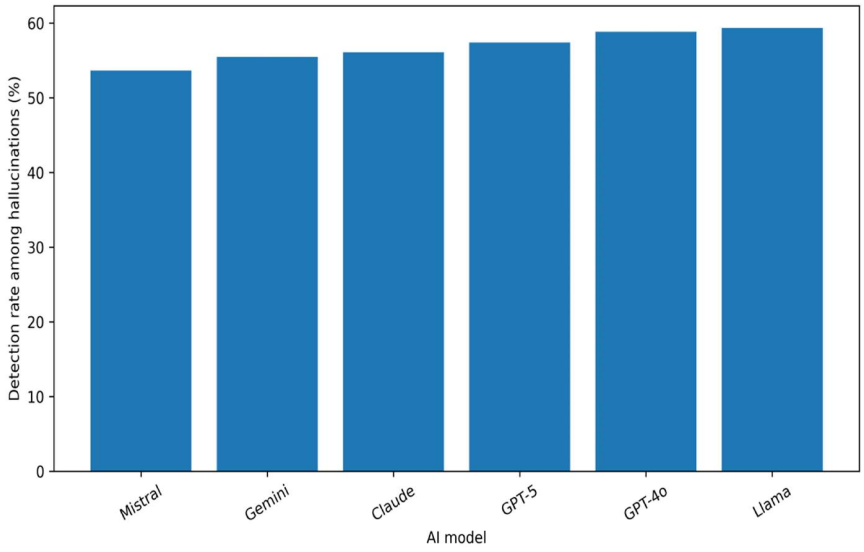

Figure 3. Hallucination detection rate by model

Figure 3 presents the proportion of hallucinated interactions that were subsequently detected for each AI model. Llama had the highest detection rate at 59.35%, followed by GPT-4o (58.86%), GPT-5 (57.42%), Claude (56.10%), Gemini (55.48%), and Mistral (53.65%).

The figure reveals an important contrast with Figure 1. Although Llama produced the highest proportion of hallucinations, it also had the highest detection rate. Conversely, Mistral produced nearly as many hallucinations as Llama but had the lowest detection rate. This distinction is important because the potential risk associated with hallucination depends not only on how frequently hallucinations occur but also on whether they are subsequently recognized.

Nevertheless, the differences in detection rate were not statistically significant. Among hallucinated interactions, the chi-square test produced $x^2(5)=4.60$, $p=0.467$, with Cramer's $V=0.039$. Therefore, interpret Figure 3 primarily as a descriptive risk pattern rather than a statistically established difference in detectability.

4.4 Hallucination Rate versus Detection Rate

Figure 4 jointly visualises hallucination frequency and hallucination detection, thereby providing a more informative representation of model-level risk than either metric considered in isolation.

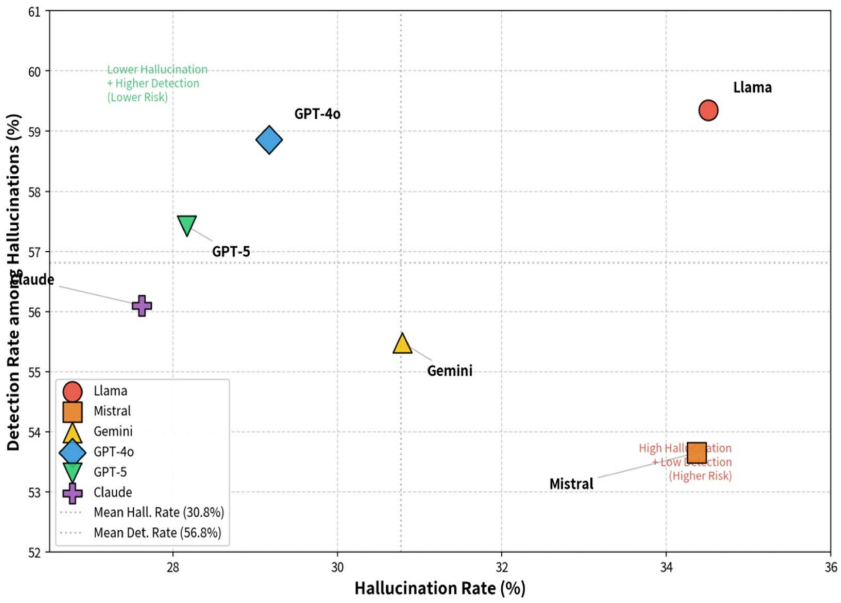

Figure 4. Hallucination Rate verus detection rate by AI Model

Llama occupies a distinctive position: it has the highest hallucination rate (34.51 %) and the highest detection rate (59.35 %). Mistral presents a contrasting and potentially more concerning configuration, combining a nearly identical hallucination rate (34.37 %) with the lowest detection rate (53.65 %). Gemini also demonstrates a relatively elevated risk combination (hallucination rate 30.79 %, detection rate 55.48 %). In contrast, GPT-5 and Claude cluster toward lower hallucination rates with intermediate to favourable detection performance. The principal interpretation is that hallucination prevalence and hallucination detectability constitute distinct dimensions of model behaviour. A high hallucination rate does not automatically imply the greatest undetected-risk exposure. Models that generate frequent hallucinations but permit relatively high detection may present lower residual risk than models that hallucinate somewhat less often yet allow a larger fraction of those errors to go unnoticed. This distinction underscores the necessity of multidimensional evaluation.

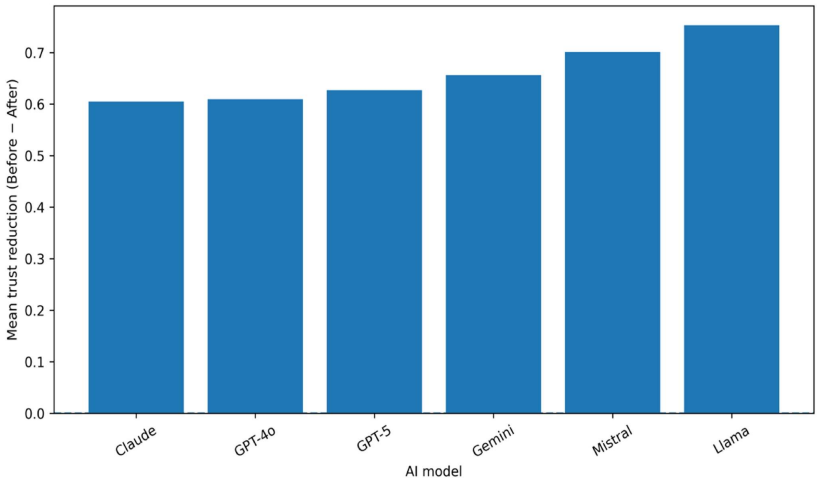

Figure 5. Mean trust reduction by model

Figure 5 shows the average reduction in user trust associated with model interactions. Across all interactions, Llama produced the largest mean trust reduction (0.75), followed by Mistral (0.70), Gemini (0.66), GPT-5 (0.63), GPT-4o (0.61), and Claude (0.60).

The figure suggests that models associated with greater hallucination exposure tended to exhibit greater overall reductions in user trust. Llama, which had the highest hallucination rate, also showed the greatest overall trust reduction. Mistral similarly combined a high hallucination rate with substantial trust reduction.

However, the relationship should not be interpreted as deterministic. GPT-4o, for example, had a lower overall trust reduction (0.61) despite having a relatively high hallucination rate (29.17%). Statistical testing showed that trust reduction across all interactions differed significantly by model (ANOVA $p=0.0001$; Welch $p=0.0001$; Kruskal–Wallis $p<0.001$), but the corresponding effect size was very small ($\omega^2=0.0021$).

Among hallucinated interactions specifically, trust reduction ranged from 1.76 for GPT-4o to 1.90 for Llama, but these differences were not statistically significant. Thus, Figure 5 indicates that model identity has some association with overall user trust reduction, but the magnitude of the model effect is limited.

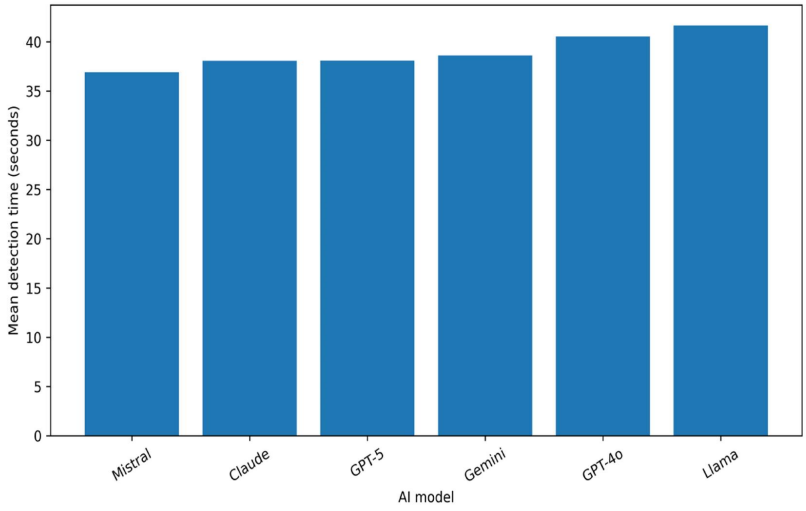

Figure 6. Mean detection time for detected hallucinations

Figure 6 presents the mean time required to detect hallucinations among interactions in which hallucinations were detected. Detection time ranged from 36.92 seconds for Mistral to 41.65 seconds for Llama. GPT-5 and Claude had mean detection times of 38.09 and 38.06 seconds, respectively, while Gemini and GPT-4o recorded 38.61 and 40.54 seconds.

The figure therefore indicates relatively modest variation in detection speed across models. Although Llama had the highest detection rate, its detected hallucinations took the longest average time to identify (41.65 seconds). Mistral showed the opposite pattern: it had the lowest detection rate but the shortest mean detection time among detected hallucinations (36.92 seconds).

This apparent contrast highlights that detection probability and detection speed are distinct dimensions. A model may have hallucinations that are relatively quickly recognized when they are detected, while still having a comparatively large proportion of hallucinations that remain undetected. Statistical testing found no significant model-level difference in detection time (ANOVA $p=0.1621$; Welch $p=0.1777$; Kruskal–Wallis $p=0.2646$; $\omega^2=0.0017$). Consequently, the differences shown in Figure 6 should be regarded primarily as descriptive.

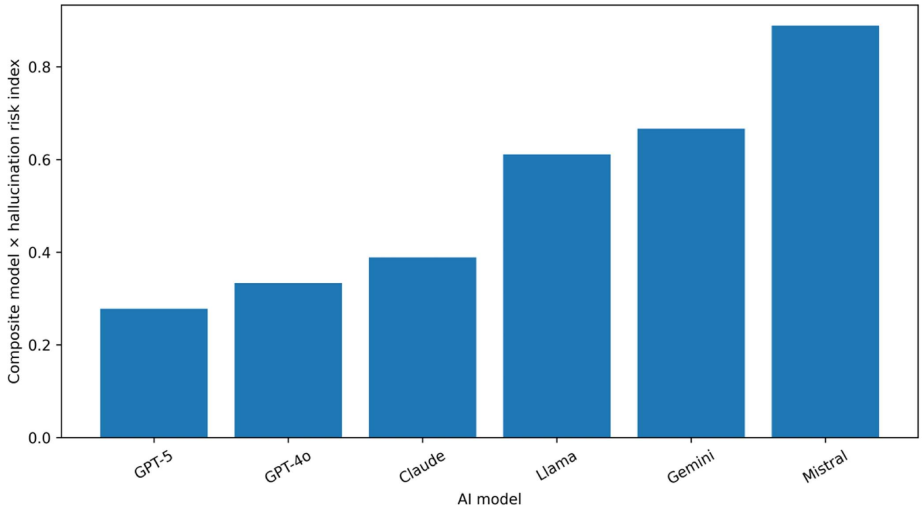

Figure 7. Exploratory model × hallucination risk index

Figure 7 presents the exploratory model × hallucination risk index, which integrates three characteristics: high hallucination rate, high AI confidence, and low hallucination detection. The resulting index was highest for Mistral (0.889), followed by Gemini (0.667), Llama (0.611), Claude (0.389), GPT-4o (0.333), and GPT-5 (0.278).

Mistral therefore occupies the most concerning position in the exploratory risk profile because it simultaneously exhibits high hallucination frequency (34.37%), very high confidence (7.997), and the lowest detection rate (53.649%). Gemini also shows a relatively elevated risk profile, whereas GPT-5 has the lowest index because of its comparatively low hallucination rate, high factual accuracy, and relatively favourable detection profile.

Importantly, the figure should not be interpreted as an absolute safety ranking. The analysis explicitly defines the index as a rank-based exploratory screening indicator rather than a validated absolute safety score. Its principal value is therefore comparative: it shows how combining multiple dimensions can reveal risk configurations that would be obscured if models were ranked solely by hallucination frequency.

5. Hallucination Synthesis

The results demonstrate that AI hallucination behaviour is multidimensional. Evaluating the model based only on hallucination frequency would overlook important differences in confidence, detection, harm, and user trust.

Mistral warrants particular attention because it combines a high hallucination rate with very high confidence and the lowest detection rate among hallucinated interactions. This configuration is potentially more consequential than hallucination frequency alone because erroneous outputs may appear highly credible while being less likely to be detected.

Llama illustrates why hallucination rate should not be interpreted in isolation. It had the highest hallucination rate but also the highest detection rate. Thus, a model with frequent hallucinations may present less undetected-risk exposure than a model with a lower hallucination rate but poorer detection.

GPT-5 showed the strongest factual accuracy profile in the dataset, combining the lowest hallucination rate with the highest factual accuracy. Nevertheless, hallucinated GPT-5 outputs were associated with substantial trust reduction, demonstrating that isolated failures can still have meaningful user consequences.

The large sample produced statistically significant differences for factual accuracy and overall hallucination occurrence even though the corresponding effect sizes were small. Accordingly, practical interpretation should focus on the joint configuration of hallucination prevalence, confidence and detection rather than statistical significance alone.

5.1 Hallucination behaviour across AI tools

The proposed simple model is

AI Tool → Hallucination → Detection → Harm →Trust

The results indicate that hallucination behaviour is broadly similar across AI tools, although some descriptive differences are visible.

1. AI Tool → Hallucination

AI Tool	N	Hallucination rate
ChatGPT	2,972	31.2%
Claude	1,576	29.1%
Gemini	2,002	30.2%
Local LLM	1,214	31.5%
Microsoft Copilot	1,013	30.1%
Perplexity	1,223	29.6%

Table 1. AI Tool and Hallucination Occurrence

Table 1 presents the distribution of hallucination rates across the six AI tools. ChatGPT accounted for 2,972 observations, with a hallucination rate of 31.2%. Claude contributed 1,576 observations and recorded a hallucination rate of 29.1%, while Gemini accounted for 2,002 observations, with a hallucination rate of 30.2%. Local LLMs represented 1,214 observations and had the highest hallucination rate among the six tools at 31.5%. Microsoft Copilot contributed 1,013 observations with a hallucination rate of 30.1%, while Perplexity contributed 1,223 observations and recorded a hallucination rate of 29.6%.

The descriptive range between the highest and lowest hallucination rates was therefore relatively narrow, extending from 29.1% for Claude to 31.5% for Local LLMs, a difference of only 2.4 percentage points. ChatGPT, Gemini, Microsoft Copilot, and Perplexity were similarly concentrated around approximately 30%. This pattern suggests that AI tools did not strongly differentiate hallucination occurrence in the dataset.

This interpretation is supported by the chi-square test, which identified no statistically significant association between AI tool and hallucination occurrence, $x^2(5) = 3.405$, $p = .638$, with a Cramer's V of .018. The extremely small effect size further indicates that the observed differences in hallucination rates have negligible practical association with AI-tool identity.

Consequently, although Local LLMs exhibited the numerically highest hallucination rate and Claude the lowest, these differences should not be interpreted as evidence that one tool is intrinsically more or less reliable than

another. Rather, the findings indicate that hallucination is a cross-tool phenomenon that occurs at broadly comparable rates across the systems examined.

5.2 Hallucination → Detection

Among the 3,039 hallucinated responses, detection rates were:

AI Tool	Hallucinations	Detected	Detection rate
ChatGPT	928	546	58.8%
Claude	458	267	58.3%
Gemini	604	339	56.1%
Local LLM	382	200	52.4%
Microsoft Copilot	305	171	56.1%
Perplexity	362	201	55.5%

Table 2. Hallucination Detection Across AI Tools

Table 2 examines what happened after hallucinations occurred by comparing the number of hallucinated responses with the number detected. Among the 3,039 hallucinated responses, ChatGPT had 928 hallucinations, of which 546 were detected, producing a detection rate of 58.8%. Claude recorded 458 hallucinations, with 267 detected, corresponding to a detection rate of 58.3%. Gemini had 604 hallucinations, of which 339 were detected, giving a detection rate of 56.1%.

Local LLMs recorded 382 hallucinations, of which 200 were detected, resulting in the lowest detection rate among the six tools at 52.4%. Microsoft Copilot had 305 hallucinations, with 171 detected, corresponding to 56.1%, while Perplexity recorded 362 hallucinations, of which 201 were detected, giving a detection rate of 55.5%.

Descriptively, ChatGPT therefore demonstrated the highest hallucination-detection rate (58.8%), closely followed by Claude (58.3%). Gemini and Microsoft Copilot both recorded 56.1%, followed by Perplexity at 55.5%, whereas Local LLMs had the lowest detection rate (52.4%). Thus, although detection performance varied by about 6.4 percentage points between the highest and lowest values, the overall pattern remained similar across tools.

The chi-square analysis confirmed that these differences were not statistically significant, $x^2(5) = 5.471$, $p = .361$, with Cramer's V = .042. The small Cramer's V indicates a very weak relationship between AI-tool identity and the probability that a hallucination would be detected.

From a reliability perspective, this result matters because it suggests that the ability to identify hallucinated outputs did not strongly depend on which AI tool generated the response. Although ChatGPT had the highest descriptive detection rate and Local LLMs the lowest, the statistical evidence does not support treating this difference as a substantive tool-level distinction. The findings instead point toward a relatively consistent limitation in hallucination detection across the evaluated systems.

5.3 Detection → Harm

Among hallucinated responses, the proportion classified as High or Critical harm was:

AI Tool	High/ Critical harm
ChatGPT	17.9%
Claude	24.0%
Gemini	20.9%
Local LLM	18.5%
Microsoft Copilot	22.2%
Perplexity	15.9%

Table 3. AI Tool and Hallucination-Related Harm

Table 3 shifts the analysis from whether hallucinations occurred and whether they were detected to the severity of their potential consequences. The table reports the proportion of hallucinated responses classified as High or critical harm for each AI tool.

Claude exhibited the highest proportion of high/critical-harm hallucinations at 24.0%. Microsoft Copilot followed with 22.2%, while Gemini recorded 20.9%. Local LLMs had a high/critical harm proportion of 18.5%, followed by ChatGPT at 17.9%. Perplexity recorded the lowest proportion at 15.9%.

The descriptive results therefore indicate some variation in the severity profile of hallucinations. In particular, the difference between Claude (24.0%) and Perplexity (15.9%) represents an 8.1-percentage-point difference in the proportion of hallucinations classified as high or critical harm. However, interpret this descriptive contrast cautiously because the overall statistical analysis found no significant differences in the distribution of harm levels across AI tools.

Specifically, the chi-square test examining the complete Low, Moderate, High, and Critical harm distribution produced $x^2(15) = 20.579$, $p = .151$, with Cramer's V = .048. The effect size again indicates only a weak association between AI-tool identity and harm severity.

Thus, Claude's higher descriptive proportion of high/critical-risk hallucinations and Perplexity's lower proportion should not be interpreted as definitive evidence that Claude generates more harmful hallucinations or that Perplexity is safer. The statistical evidence does not establish a significant tool-specific difference. Instead, the results indicate that potentially harmful hallucinations represent a meaningful reliability concern across the AI tools considered.

5.4 Hallucination → Trust

This is the strongest relationship in the analysis.

Across all observations:

- No hallucination: mean trust change = -0.140

• Hallucination: mean trust change = -1.823

Therefore, hallucinations are associated with a much larger reduction in user trust.

Interestingly, trust loss following hallucination was quite similar across tools:

AI Tool	Mean trust change after hallucination
ChatGPT	-1.809
Claude	-1.774
Gemini	-1.895
Local LLM	-1.815
Microsoft Copilot	-1.821
Perplexity	-1.809

Table 4. AI Tool and Trust Change Following Hallucination

Table 4 presents the relationship between hallucination occurrence and subsequent changes in user trust. This analysis provides the strongest substantive finding across the four tables. Across all observations, responses without hallucination were associated with a mean trust change of “0.140, whereas hallucinated responses were associated with a much larger mean trust change of -1.823.

This difference demonstrates that hallucination occurrence was associated with a substantially greater deterioration in user trust than non-hallucinatory responses. The magnitude of the difference approximately 1.683 trust units is considerably larger than the relatively small differences observed among AI tools. This indicates that whether a hallucination occurred mattered far more for trust than which AI tool produced it. The tool specific results reinforce this interpretation. ChatGPT had a mean trust change of “1.809 following hallucination. Claude recorded -1.774, Gemini -1.895, Local LLM -1.815, Microsoft Copilot “1.821, and Perplexity -1.809.

The values cluster around -1.8, with Gemini showing the largest trust reduction (-1.895) and Claude the smallest (-1.774). The difference between these extremes is only 0.121 trust units, demonstrating the remarkable similarity of trust deterioration across tools.

An ANOVA confirmed that mean trust change did not differ significantly by AI tool, $F(5, 9994) = 0.919$, $p = .467$. Therefore, the evidence does not support the conclusion that users respond substantially differently to hallucinations depending on which AI system generated them.

An additional finding concerns whether detecting a hallucination mitigated the associated trust penalty. Among hallucinated responses, mean trust change was -1.804 when the hallucination was detected, compared with -1.848 when it was not detected. The difference was therefore only 0.044 trust units, suggesting that detection alone produced little apparent restoration of confidence. This finding is particularly important because it implies that identifying an erroneous AI response may not be sufficient to reverse the damage to user confidence

once the hallucination has occurred

6. Discussion

The above Tables 1–4 reveal a consistent pattern across the proposed AI Tool → Hallucination → Detection → Harm →Trust pathway. First, hallucination rates were broadly comparable across the six AI tools, ranging from 29.1% to 31.5%, and the association between AI tool and hallucination occurrence was negligible (Cramer's V = .018). Second, detection rates also showed only limited variation, ranging from 52.4% to 58.8%, with a very small association between tool identity and detection (Cramer's V = .042).

Third, although the proportion of high/critical harm hallucinations varied descriptively from 15.9% to 24.0%, the complete harm distribution did not differ significantly between tools (Cramer's V = .048). Finally, trust deterioration following hallucination was remarkably consistent across all six tools, with mean trust changes ranging from -1.774 to -1.895, and the ANOVA showing no significant tool level effect (p = .467).

The principal finding is therefore not that one AI tool is demonstrably more reliable than another, but rather that hallucination itself represents a cross-tool reliability problem with a substantial consequence for user trust. The analysis shows a clear contrast between the weak explanatory role of AI-tool identity and the strong practical association between hallucination occurrence and trust deterioration.

Accordingly, the findings caution against ranking the six AI tools as categorically more or less reliable based solely on these data. Instead, the results support a broader interpretation in which hallucination risk, imperfect detection, potential harm, and trust degradation constitute system level challenges shared across contemporary AI tools. The particularly small difference in trust loss between detected and undetected hallucinations further suggests that detection does not necessarily eliminate the negative trust consequence of an erroneous response.

Overall, the four tables provide empirical support for conceptualizing hallucination as a cross platform reliability and human AI trust problem, rather than primarily as a problem attributable to a particular AI tool. This distinction is important for the subsequent modeling of the proposed pathway because it suggests that hallucination occurrence and its downstream consequences may have considerably greater explanatory value than AI-tool identity itself. The analysis therefore provides a strong rationale for further mediation/path analysis, logistic or ordinal regression, and trust-prediction models that control for baseline trust and hallucination characteristics.

The key conclusion would be that tool identity has weak explanatory power, whereas hallucination occurrence has a strong practical relationship with user trust.

7. Conclusion

This study examined hallucination related risk across six contemporary AI models using a large interaction dataset (N = 10,000). By jointly analysing hallucination frequency, confidence accuracy relationships, detection performance, harm potential, and trust impact, the work advances beyond single metric assessments of model reliability and offers an integrated view of multidimensional risk.

7.1 Principal Empirical Findings

Hallucination occurrence differed significantly across models ($x^2(5) = 32.68$, $p < .001$), yet the association was small (Cramer's V = 0.057). Llama (34.51 %) and Mistral (34.37 %) exhibited the highest rates, while Claude (27.62 %) and GPT-5 (28.17 %) recorded the lowest. Factual accuracy also differed reliably across models, with GPT-5 achieving the highest mean accuracy and Mistral the lowest; however, effect sizes remained modest ($\omega^2 \approx 0.003$). In contrast, mean AI confidence was uniformly high (approximately 7.9–8.0) and did not differ significantly across models. This pattern produces a clear confidence accuracy mismatch: models with comparatively lower factual accuracy did not correspondingly lower their expressed confidence.

Detection rates among hallucinated interactions ranged from 53.65 % (Mistral) to 59.35 % (Llama) but did not differ significantly. Detection times were likewise statistically indistinguishable across models. Harm severity distributions and verification rates showed no reliable model-level differentiation. Trust reduction across all interactions did differ significantly by model, yet the effect size was again very small ($\omega^2 = 0.0021$). Among hallucinated interactions specifically, mean trust reduction was substantial (approximately 1.8 scale points) and did not vary meaningfully by model.

7.2 Risk Configuration Insights

The most consequential insight comes from examining frequency, confidence, and detectability together. Mistral presented the clearest high risk configuration: a high hallucination rate, the highest mean confidence, and the lowest detection rate. Llama, despite matching Mistral's high hallucination frequency, benefited from the highest detection rate and therefore occupied a less severe residual risk position. GPT-5 combined the lowest hallucination rate with the highest factual accuracy and a comparatively favourable detection profile, yielding the lowest exploratory risk index score. These configurations demonstrate that ranking models solely by hallucination rate would misrepresent relative safety; detectability and confidence calibration must be considered simultaneously.

7.3 Cross-Tool Consistency and the Primacy of Hallucination Itself

When interactions were aggregated by AI tool rather than by specific model, hallucination rates, detection rates, harm distributions, and post hallucination trust changes showed even weaker differentiation. Tool identity explained negligible variance in these outcomes. By contrast, the simple presence versus absence of a hallucination produced a large and consistent trust penalty (mean trust change $\approx$-0.14 without hallucination versus $\approx$-1.82 with hallucination). Detecting the hallucination mitigated this penalty only marginally (a difference of about 0.04 trust units). These results indicate that hallucination is a systemic, cross platform reliability problem whose trust consequences are driven primarily by the occurrence of the error rather than by the identity of the generating system.

7.4 Theoretical and Practical Implications

Theoretically, the findings reinforce the view that hallucination should be conceptualised as a multidimensional construct encompassing generation frequency, epistemic calibration (confidence versus accuracy), detectability, and downstream human consequences. Methodologically, they highlight the limits of isolated factuality metrics and support using composite risk profiles for comparative evaluation.

Practically, the results caution against treating any single model as categorically safe or unsafe based on hallucination rate alone. Organisations deploying LLMs in high stakes settings should prioritise (a) monitoring of confidence accuracy alignment, (b) investment in detection and verification mechanisms that operate effectively across models, and (c) user interface designs that encourage appropriate scepticism even when outputs appear fluent and confident. Because detection itself does little to restore lost trust, prevention and early interception of hallucinations remain preferable to post hoc correction.

7.5 Future Directions

Several limitations should be acknowledged. The analysis is observational and relies on the source dataset's operational definitions and coding schemes; causal claims about model architecture or training procedures cannot be made. Effect sizes for model differences were consistently small, indicating that practical differences, while detectable in a large sample, may be modest in absolute terms. The exploratory risk index is rank-based and has not been externally validated. Future work should incorporate controlled experimental designs, longitudinal trust measurement, domain specific harm taxonomies, and multimodal systems, as well as formal mediation and path models that quantify the relative contributions of frequency, confidence, detection, and harm to trust outcomes.

7.6 Closing Statement

In summary, AI hallucination risk is multidimensional. Frequency, confidence accuracy mismatch, detectability, harm potential, and trust impact each contribute distinct information about system reliability.

Among the models examined, Mistral exhibited the most concerning joint profile, while GPT-5 showed the most favourable accuracy and frequency characteristics. Across tools, hallucination itself, not tool identity, emerged as the dominant driver of trust erosion. Improving the trustworthiness of AI systems therefore requires coordinated attention to all four dimensions of risk rather than exclusive focus on reducing the raw rate of factual errors.

References

[1] Kabir, Z. S. (2026). Explaining hallucination of artificial intelligence in data driven decision-making. In *Innovative decision making in engineering: The role of cognition, heuristics, and human factors* (pp. 85–112). IGI Global Scientific Publishing.

[2] Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, R., Rawte, V., Chadha, A., Das, A. (2024). A compreh ensive survey of hallucination mitigation techniques in large language models. *arXiv*. https://arxiv.org/abs/2401.01313

[3] Rawte, V., Sheth, A., Das, A. (2023). A survey of hallucination in large foundation models. *arXiv*. https://arxiv.org/abs/2309.05922.

[4] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T. (2024). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*. https://doi.org/10.1145/3703155.

[5] Xu, Z., Jain, S., Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. *arXiv*. https://arxiv.org/abs/2401.11817.

[6] Kalai, A. T., Vempala, S. S. (2023). Calibrated language models must hallucinate. *arXiv*. https://arxiv.org/abs/2311.14648.

[7] Hassan, M. (2025). Measuring the impact of hallucinations on human reliance in LLM applications. *Journal of Robotic Process Automation, AI Integration, and Workflow Optimization, 10*(1), 10–20.

[8] Yin, Z. (2024). A review of methods for alleviating hallucination issues in large language models. *Applied and Computational Engineering, 76*(1), 258–266. https://doi.org/10.54254/2755-2721/76/20240608.

[9] Rawte, V., Sheth, A., Das, A. (2023). A survey of hallucination in large foundation models. *arXiv*. https://arxiv.org/abs/2309.05922 *(Duplicate of #3)*

[10] Luo, J., Li, T., Wu, D., Jenkin, M., Liu, S., Dudek, G. (2024). Hallucination detection and hallucination mitigation: An investigation. *arXiv*. https://arxiv.org/abs/2401.08358.

[11] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., Fung, P. (2024). Survey of hallucination in natural language generation. *arXiv*. https://arxiv.org/abs/2202.03629.

[12] Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., Tonmoy, S. M. T. I., Chadha, A., Sheth, A., Das, A. (2023). The troubling emergence of hallucination in large language models: An extensive definition, quantif- ication, and prescriptive remediations. *arXiv*. https://arxiv.org/abs/2310.04988.

[13] Liu, F., Liu, Y., Shi, L., Huang, H., Wang, Y., Zhang, Z., Sun, J. (2024). Exploring and evaluating hallucinations in LLM powered code generation. *arXiv*. https://arxiv.org/abs/2404.00971.

[14] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., Shi, S. (2023). Siren's song in the AI ocean: A survey on hallucination in large language

models. *arXiv*. https://arxiv.org/abs/2309.01219.

[15] Perkoviæ, G., Drobnjak, A., Botièki, I. (2024). Hallucinations in LLMs: Understanding and addressing challenges. In *2024 47th ICT and Electronics Convention (MIPRO)* (p. 2084–2088). IEEE. https://doi.org/10.1109/MIPRO60963.2024.10569238.

[16] Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Esmaeilpour, N., Ganguli, D., Hume, D., Keeling, S., Kaplan, J. (2023). Towards understanding sycophancy in language models. *arXiv*. https://arxiv.org/abs/2310.13548.

[17] Zhao, Y., Liu, Z., Zheng, Y., Lam, K.-Y. (2025). Attribution techniques for mitigating hallucination in RAG-based question-answering systems: A survey. *TechRxiv*. https://doi.org/10.36227/techrxiv.175036754.46793269/v1 *(Note: Your file lists this as #17 but also has Wei et al. listed as #17—duplicate number)*

[18] Zhang, W., Zhang, J. (2025). Hallucination mitigation for retrieval-augmented large language models: A review. *Mathematics, 13*(5), Article 856. https://doi.org/10.3390/math13050856.

[19] Li, Y., Li, Z., Hung, K., Wang, W., Xie, H., Li, Y. (2025). Ambiguity processing in large language models: Detection, resolution, and the path to hallucination. *Natural Language Processing Journal*, Article 10- 0173. https://doi.org/10.1016/j.nlp.2025.100173.

[20] Agrawal, G., Kumarage, T., Alghamdi, Z., Liu, H. (2024). Can knowledge graphs reduce hallucinations in LLMs A survey. *arXiv*. https://arxiv.org/abs/2311.07914.

[21] Bilal, A., Mohsin, M. A., Umer, M., Bangash, M. A. K., Jamshed, M. A. (2025). Meta-thinking in LLMs via multi-agent reinforcement learning: A survey. *arXiv*. https://arxiv.org/abs/2504.14520.

[22] Kazlaris, I., Antoniou, E., Diamantaras, K., Bratsas, C. (2025). From illusion to insight: A taxonomic survey of hallucination mitigation techniques in LLMs. *AI, 6*(10), 260.

[23] Jha, S., Jha, S. K., Lincoln, P., Bastian, N. D., Velasquez, A., Neema, S. (2023). Dehallucinating large language models using formal methods guided iterative prompting. In *Proceedings of the 2023 IEEE International Conference on Assured Autonomy (ICAA)* (p. 149–152). IEEE. https://doi.org/10.1109/ICAA58325.2023.00029.

[24] Turpin, M., Michael, J., Perez, E., Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv*. https://arxiv.org/abs/2305.04388.

[25] Si, C., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J., Wang, L. (2022). Prompting GPT-3 to be reliable. *arXiv*. https://arxiv.org/abs/2210.09150.

[26] Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. https://doi.org/10.1145/3544548.3581388.

[27] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv*. https://arxiv.org/abs/2201.11903 *(Your file has #17 assigned to this but corrected to #27 for order).*

[28] Jha, S., Jha, S. K., Lincoln, P., Bastian, N. D., Velasquez, A., Neema, S. (2023). Dehallucinating large language models using formal methods guided iterative prompting. In *Proceedings of the 2023 IEEE International Conference on Assured Autonomy (ICAA)* (p. 149–152). IEEE. https://doi.org/10.1109/ICAA58325.2023.00029 *(Duplicate of #23)*

[29] Mündler, N., He, J., Jenko, S., Vechev, M. (2023). Self contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv*. https://arxiv.org/abs/2305.15852.

[30] Azaria, A., Mitchell, T. (2023). The internal state of an LLM knows when it's lying. *arXiv*. https://arxiv.org/abs/2304.13734.

[31] Luo, J., Xiao, C., Ma, F. (2023). Zero resource hallucination prevention for large language models. *arXiv*. https://arxiv.org/abs/2309.02654.

[32] Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Zhou, D. (2023). FreshLLMs: Refreshing large language models with search engine augmentation. *arXiv*. https://arxiv.org/abs/2310.03214.

[33] Lai, K. E., Mustaffa, N., Preece, C. N., Wong, F. Y. (2026). Hallucinations in large language models within academic contexts: A systematic review. *Journal of Science and Technology Policy Management*, 1–28.

[34] Liu, Z., Wang, Y., Chen, H., Zhang, Q., Li, S., Li, Y. (2024). Comprehensive evaluation of AI hallucination and novel UV-oriented framework toward safe and trustworthy AI. In *2024 7th International Conference on Universal Village (UV)* (p. 1–136). *IEEE*.

[35] Banerjee, S., Agarwal, A., Singla, S. (2024). LLMs will always hallucinate, and we need to live with this. *arXiv*. https://doi.org/10.48550/arXiv.2409.05746