Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When AI Fails: Hallucination Exposure, Verification, and Critical AI Literacy

Lukasz Nowacki

lukasz.nowacki2@wsb.edu.pl

WSB University

Agata Ewa Wrochna

Wenzhou-Kean University

Research Article

Keywords: Generative AI, AI hallucinations, Critical AI Literacy, external verification, trust calibration, epistemic risk

Posted Date: October 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11132351/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

The rapid integration of generative AI into everyday information practices has created a central contradiction: its responses may appear coherent, confident, and authoritative while remaining inaccurate or entirely fabricated. This study examines how users experience and respond to AI hallucinations and what these responses reveal about the development of Critical AI Literacy. Drawing on survey data from 314 respondents, the analysis explores relationships between hallucination exposure, conceptual awareness, perceived AI opacity and severity, knowledge of AI limitations, and verification and reliance practices. The findings reveal a marked gap between experience and understanding. While 82.3% of respondents had encountered inaccurate or fabricated AI-generated information at least occasionally, only 42.6% were familiar with the concept of AI hallucinations. Most users nevertheless did not reject AI after experiencing such failures. Instead, they checked information against external sources, used AI in an auxiliary role, or employed control prompts intended to reduce unreliable outputs. When the relevant factors were considered together, exposure to hallucinations was the only factor that predicted external verification; general AI experience and frequency of use failed to do so. This suggests that experience of AI is not equivalent to experience of AI failure. The study therefore conceptualises Critical AI Literacy as a form of epistemic adaptation that extends beyond technical competence. It involves recognising that fluency does not guarantee accuracy, maintaining an appropriate degree of critical distance, and adjusting reliance according to context and the possibility of independent verification.

INTRODUCTION

The rapid development of generative artificial intelligence (AI) has significantly changed the way information is produced, accessed, and evaluated across education, professional environments, and everyday communication. Large language models (LLMs), such as ChatGPT and similar systems, are able to generate coherent, contextually appropriate, and often highly convincing textual outputs (Brown et al., 2020; OpenAI, 2023). At the same time, one of their most problematic characteristics is the tendency to produce information that is factually incorrect, misleading, or entirely fabricated, commonly referred to as “AI hallucinations” (Ji et al., 2023; Maynez et al., 2020). These outputs are increasingly regarded not simply as occasional technical errors, but as a structural consequence of probabilistic language generation, where linguistic plausibility may take precedence over factual accuracy (Bender et al., 2021; Ooi et al., 2023). This creates a particular problem for users. Unlike traditional information sources, where credibility can often be assessed through identifiable authors, references, or institutional authority, generative AI may produce information that appears reliable without providing clear evidence for the claims it makes. Users are therefore presented with outputs that can be convincing and useful while remaining potentially inaccurate (Monteith et al., 2024). In education, this raises concerns about the reliability of AI-assisted learning, the quality of knowledge construction, and users’ ability to critically assess AI-generated content (Kasneci et al., 2023; Zawacki-Richter et al., 2019). Existing research has approached these challenges through the broader concepts of digital literacy and AI literacy, focusing on competencies such as understanding AI systems, formulating effective prompts, evaluating outputs, and

integrating AI into learning and professional practices (Ng et al., 2021; Long & Magerko, 2020). These skills are important; however, knowing how to use an AI system efficiently does not necessarily mean being able to recognise when its output is unreliable. The growing use of generative AI therefore requires users not only to operate the technology but also to judge the credibility and factual grounding of the information it provides.

Building on this distinction, the present study develops the concept of Critical AI Literacy as an epistemic response to the unreliability of generative AI. Drawing on a large-scale survey of users' experiences, awareness, and perceptions of AI hallucinations, the study examines how individuals recognise and respond to inaccurate or fabricated AI-generated information. The findings indicate that many users do not simply accept or reject AI outputs. Instead, they check information against external sources, use AI in an auxiliary role, or attempt to reduce the possibility of hallucinations through control prompts (Metzger and Flanagin, 2015; Ou and Ho, 2024). Such practices suggest a form of literacy based not only on knowing how to use AI, but also on understanding its limitations and adjusting reliance accordingly. This process is complicated by the perceived opacity of generative AI systems. Users may receive detailed and confident responses while having limited understanding of how they were produced or how reliable they are (Shin et al., 2024). Under such conditions, trust needs to be adjusted according to the context, the perceived reliability of the output, and the possibility of independent verification. At the same time, responsibility for dealing with hallucinations cannot rest solely with individual users. Although users may develop strategies aimed at reducing the risks associated with inaccurate AI-generated information, developers, institutions, and regulatory bodies also shape the conditions under which these systems are used. Critical AI Literacy should therefore not be treated as a substitute for system transparency, responsible design, or appropriate governance. Rather, it forms one part of a broader response to the epistemic and communicative risks created by generative AI (Mittelstadt et al., 2016; Floridi et al., 2018). Against this background, the present study asks how users experience and conceptualise AI hallucinations, how they evaluate the reliability of AI-generated information, and what strategies they employ to manage the uncertainty associated with such outputs. By addressing these questions, the paper contributes to the developing discussion on AI literacy by placing greater emphasis on its epistemic dimension: the capacity to recognise unreliable information, verify questionable outputs, and make informed decisions about when AI can and cannot be relied upon.

In this sense, AI hallucinations represent more than a technical limitation of contemporary language models. They expose a broader shift in the conditions under which information is produced and evaluated. As generative AI increasingly enters education, professional practice, and everyday information seeking, users are required to distinguish between language that merely appears authoritative and information that can be independently supported. The concept of Critical AI Literacy as epistemic adaptation proposed in this study is intended to capture this changing relationship between users, AI-generated information, and the question of what should be accepted as reliable knowledge.

LITERATURE REVIEW

Hallucinations and the Emergence of Epistemic Risk

Hallucinations are commonly understood as outputs that are factually incorrect, fabricated, or inconsistent with external reality (Ji et al., 2023). The phenomenon, however, extends beyond isolated factual errors. Sun et al. (2024) identify different forms of hallucination, including factual fabrication, reasoning failures, logical inconsistencies, mathematical errors, and contextual distortions. Liu and Lin (2026) similarly argue that hallucinations should not simply be treated as software bugs but as communicative failures resulting from the probabilistic nature of language generation. LLMs are designed to produce plausible linguistic responses rather than independently establish whether those responses represent verified knowledge. Hallucinations can therefore be understood as a broader form of epistemic distortion arising from the way generative systems produce information. Recent scholarship also suggests that hallucinations are structural rather than incidental. Kalai et al. (2026) and Banerjee et al. (2025) argue that despite developments in retrieval-augmented generation, reasoning modules, and post-training alignment, hallucinations remain an inherent consequence of probabilistic language generation. Such mechanisms may reduce their frequency but cannot entirely eliminate the underlying problem. This distinction becomes particularly important as AI systems improve. Li et al. (2026) describe an accuracy paradox, whereby increasingly accurate systems may strengthen users' confidence while reducing their willingness to question or independently verify AI-generated information. The epistemic risk associated with hallucinations therefore depends not only on how often inaccurate information is produced, but also on whether users are able to recognise when further verification is necessary. This problem can also be considered in relation to classical epistemology. Popper (1959) emphasised systematic attempts at falsification in the development of knowledge, while Goldman (1986) focused on the reliability of the processes through which beliefs are formed. Generative AI complicates these assumptions by producing responses that display characteristics commonly associated with expertise, fluency, coherence, and confidence, without necessarily satisfying the epistemic conditions associated with reliable knowledge. Users may consequently encounter information that appears trustworthy while having limited access to the evidence on which it is based.

The implications may extend beyond individual interactions with AI. Obiefuna (2025) argues that as AI-generated content increasingly circulates online and becomes incorporated into future training data, fabricated information may become more difficult to distinguish from verified knowledge. He describes this recursive process as epistemic collapse, pointing to the possibility that hallucinations may affect not only individual judgments but also the reliability of wider digital knowledge environments. Hallucinations should therefore be understood as more than isolated AI errors. They reveal a growing tension between linguistic plausibility and epistemic justification, creating an information environment in which convincing language does not necessarily represent trustworthy knowledge. This raises a further question concerning how users decide what and whom to trust when interacting with generative AI.

Trust, Authority and the Illusion of Knowing

Knowledge has long been recognised as a social process. Individuals cannot independently verify every piece of information they encounter and consequently rely on experts, institutions, and other sources of epistemic authority Trust plays an important role in this process. Luhmann (1979, 1988) argues that trust reduces complexity by allowing individuals to make decisions under conditions of uncertainty, while Giddens (1990) emphasises the growing dependence of modern societies on abstract systems of expertise. More recently, Marková (2025) defines epistemic trust as the willingness to accept information from others because they are recognised as possessing greater knowledge or expertise. Trust is therefore an important mechanism through which knowledge is acquired and accepted.

Generative AI complicates this relationship between trust and authority. Although LLMs possess neither beliefs nor genuine understanding, they communicate in ways that resemble knowledgeable human actors. Their responses are often fluent, coherent, confident, and immediate, characteristics that may encourage users to interpret the information they provide as credible. This is consistent with research showing that people frequently use fluency and ease of processing as indicators of truth or reliability (Tversky & Kahneman, 1974; Kahneman, 2011). The illusory truth effect similarly demonstrates that information that is repeated or easier to process may be perceived as more credible regardless of its factual accuracy (Fazio et al., 2015). In generative AI, linguistic confidence and coherence can therefore be mistaken for epistemic reliability. The problem extends beyond the credibility of individual responses. Messeri and Crockett (2024) argue that AI systems may create illusions of understanding by providing coherent explanations that increase users' confidence while reducing some of the cognitive effort normally required to acquire knowledge. This resembles the illusion of explanatory depth identified by Rozenblit and Keil (2002), whereby individuals overestimate how well they understand complex phenomena. Nowotny (2024) identifies a related illusion of control associated with predictive technologies, where algorithmic systems may create the impression that uncertainty has been replaced by computational certainty. Users may consequently overestimate not only the reliability of AI-generated information but also their own understanding of the subjects on which they rely on AI. Generative AI may therefore influence how epistemic authority itself is established. Palese (2026) argues that AI systems increasingly function as epistemic infrastructures, influencing what users recognise as credible and trustworthy. Unlike traditional forms of expertise, however, the authority of generative AI is not necessarily based on identifiable knowledge, qualifications, or verifiable evidence. It may instead emerge from the system's ability to provide immediate and apparently knowledgeable responses. Quattrociocchi et al. (2025) similarly emphasise that LLMs generate statistically plausible language rather than epistemically grounded knowledge, while Li et al. (2026) demonstrate that improvements in factual accuracy do not necessarily resolve the epistemic risks associated with growing confidence in these systems. The problem is therefore not simply whether AI provides correct information, but how users determine when its apparent authority is justified. Reliance on AI is also shaped by users' ability to assess their own knowledge and the capabilities of the system. He et al. (2023) show that appropriate reliance depends partly on users' capacity to evaluate both accurately. Metacognitive biases, including the Dunning–Kruger effect (Dunning 2011), may lead users to misjudge their own competence and consequently miscalibrate their reliance on AI. Their findings also indicate that interventions aimed at

improving self-assessment can contribute to better trust calibration. Effective human-AI interaction therefore does not require users to maximise or eliminate trust, but to judge when reliance is justified and when additional scrutiny is necessary.

Verification Practices and Trust Calibration

Recognition of AI unreliability does not necessarily result in rejection of the technology. Users may instead adjust the degree of trust they place in AI-generated information and adopt strategies intended to reduce the risk of relying on inaccurate outputs. This process can be understood through the concept of trust calibration, which refers to the alignment between users' reliance on a system and its actual capabilities and limitations (Okamura & Yamada, 2020). From this perspective, the objective is not to maximise trust in AI but to develop an appropriate level of reliance according to the context and potential consequences of error.

Trust calibration is increasingly recognised as an important component of human-AI interaction. Liebherr et al. (2026) argue that calibrated trust develops through continued interaction as users encounter both the capabilities and limitations of AI systems. Duan & Zhao (2024) similarly distinguish between initial trust, situational trust, and learned trust, suggesting that reliance changes as users accumulate experience with a system. This is particularly relevant to generative AI, where users may initially be influenced by fluent and confident outputs but subsequently revise their expectations after encountering inaccurate or fabricated information. One practical expression of this adjustment is verification. Rather than accepting AI-generated information at face value, users may compare outputs with independent sources, seek supporting evidence, or limit AI to an auxiliary role in decision-making. Verification therefore introduces an additional layer of evaluation between the production of an AI-generated response and its acceptance as reliable information. Verification can also be understood as a form of epistemic vigilance. Sperber et al. (2010) use this concept to describe the mechanisms through which individuals assess communicated information before accepting it as reliable. Such vigilance is particularly relevant to generative AI because conventional indicators of credibility, such as confidence, coherence, and apparent expertise, may not reliably indicate factual accuracy. External verification allows users to compensate, at least partly, for this uncertainty by comparing AI-generated claims with sources whose epistemic basis can be more readily established. Importantly, verification does not imply complete distrust of AI. Excessive scepticism may reduce the benefits users can obtain from systems that are often useful and accurate, just as excessive trust increases the risk of inappropriate reliance. He et al. (2023) show that users' ability to assess both their own knowledge and AI capabilities plays an important role in this process. Verification may form part of this calibration by allowing users to retain the practical benefits of generative AI while reducing the epistemic risks associated with accepting its outputs uncritically.

This perspective suggests that encounters with AI failure may play an important role in shaping subsequent behaviour. If users recognise that apparently credible outputs can be inaccurate, they may become more likely to verify information or restrict the contexts in which they rely on AI.

Critical AI Literacy as Epistemic Adaptation

The increasing use of generative AI has made it clear that technical proficiency alone is insufficient for effective interaction with these systems. Existing AI literacy frameworks generally define AI literacy as a combination of knowledge, skills, and attitudes that enable individuals to understand AI technologies, evaluate their capabilities and limitations, and use them responsibly (Long & Magerko, 2020; Ng et al., 2021). These frameworks have made an important contribution by moving the discussion beyond technical expertise towards broader competencies such as critical evaluation, ethical awareness, and informed decision-making.

This perspective is particularly important in education, where generative AI is increasingly used to search for information, explain concepts, support writing, and assist with learning. Ismaili (2024) argues that AI literacy should be integrated into curricula alongside critical thinking, while Danyaro et al. (2025) suggest that the educational challenges associated with LLM hallucinations cannot be addressed solely through technical solutions such as prompt engineering or retrieval-augmented generation. Students also need to recognise uncertainty, question AI-generated outputs, and critically evaluate the information these systems provide. Hallucinations, however, introduce a particular challenge for AI literacy. Existing frameworks already recognise the importance of critical evaluation, but generative AI confronts users with information that may be simultaneously plausible and unreliable. As discussed in the previous sections, users cannot necessarily determine the reliability of an output from the way it is presented. Knowing how AI works or how to formulate an effective prompt is therefore only part of the problem. Users must also be able to determine when an apparently credible answer requires additional scrutiny and when reliance on AI should be limited. The concept of Critical AI Literacy provides a useful basis for addressing this problem. Strauß (2021) argues that the widespread adoption of AI requires greater awareness of automation bias, uncertainty, and the potential mismatch between system behaviour and users' expectations. Ferreira (2026) as well as Tanchuk (2026) similarly argue that the epistemic risks associated with AI require stronger epistemic literacy and self-regulation among learners, including a realistic understanding of the limits of AI-generated knowledge. Broader developments in digital literacy point in the same direction. Yao and Lin (2026) emphasise the need to move beyond technical competencies towards contextual, cultural, and critical dimensions of AI use, while Tsao et al. (2025) highlight the importance of recognising the contextual and situated character of knowledge in human–AI collaboration.

Building on these perspectives, the present study conceptualises Critical AI Literacy as a form of epistemic adaptation to hallucinating systems (Fig. 1.). The concept does not replace existing approaches to AI literacy or represent an entirely separate set of competencies. Rather, it places greater emphasis on how users respond to the uncertainty created by systems capable of producing convincing but potentially inaccurate information. From this perspective, Critical AI Literacy involves recognising the limits of AI-generated knowledge, questioning outputs when their reliability cannot be established, verifying information through independent sources, and adjusting reliance according to context and the potential consequences of error.

The notion of epistemic adaptation shifts attention from literacy as a relatively static set of skills towards the changing relationship between users and generative AI. Encounters with unreliable outputs may lead users to reconsider the authority they grant these systems and adopt more cautious forms of reliance. Verification and trust calibration can therefore be understood as possible expressions of this process. Critical AI Literacy, in this sense, does not require users to reject generative AI or approach all AI-generated information with suspicion. Instead, it involves learning to use these systems while recognising that linguistic fluency and apparent confidence do not guarantee factual reliability. This conceptualisation provides the basis for examining how experiences of AI hallucinations relate to perceptions of opacity and risk, as well as to the verification and reliance strategies explored in the empirical analysis.

METHODOLOGY

Research Design

The study adopted a survey-based research design to examine how users respond to AI hallucinations and whether encounters with unreliable AI-generated content are associated with perceptions of AI opacity and epistemic risk, awareness of AI limitations, and verification practices. A predominantly quantitative approach was employed, with open-ended responses providing additional context for interpreting selected findings. The analysis focused on five dimensions directly related to the proposed theoretical framework, whose operationalisation is presented in Table 1.

Table 1. Operationalisation of the main constructs

Construct	Operationalisation	Questionnaire items
Experience with AI hallucinations	Awareness of AI hallucinations, frequency of encounters with inaccurate or fabricated AI-generated information, and perceived scale of the problem	Q16–20
Perceived AI opacity	Perceived transparency, rationality, and adequacy of AI explanations	Q9–15
Verification and reliance strategies	External verification, control prompting, auxiliary/supporting use of AI, and absence of mitigation strategies	Q23
Knowledge of AI limitations	Self-reported knowledge of the limitations and risks associated with generative AI	Q26
Critical AI Literacy	Theoretical interpretation synthesised from relationships among experience, perceptions of opacity and risk, knowledge of AI limitations, and reported verification and reliance practices	—

The study was conducted in February 2026 and was informed by a communication-epistemic perspective, treating AI hallucinations not simply as technical failures but as communicative phenomena that affect how users evaluate the credibility, reliability, and authority of AI-generated information. Critical AI Literacy was not measured as a separate scale but interpreted theoretically through the relationships between users' experiences, perceptions, knowledge of AI limitations, and reported verification practices.

Sample and Survey Instrument

The study involved 314 respondents recruited using non-probability convenience sampling. The online questionnaire was distributed through digital platforms frequented by users of generative AI, reaching participants with varying levels of experience with tools such as ChatGPT, Copilot, Gemini, and other large language models. Participation was voluntary and anonymous. The sample included Polish university students, educators, professionals from different industries, and individuals using generative AI for personal purposes. Respondents also varied considerably in age and experience with generative AI (Table 2), ranging from no previous experience to advanced and expert users.

Given the exploratory nature of the study, the sample was not intended to be representative of the general population. Instead, the sampling strategy was designed to capture a broad range of experiences and patterns of interaction with generative AI. This enabled the examination of relationships between exposure to AI hallucinations, perceptions of AI opacity and risk, knowledge of AI limitations, and verification practices.

Table 2. Demographic profile of the sample

Gender			**Experience using generative AI**		
Category	***n***	**%**	**Category**	***n***	**%**
Female	192	61.1	None	23	7.3
Male	119	37.9	Basic	175	55.7
No response	3	1.0	Intermediate	82	26.1
Age			Advanced	23	7.3
Category	***n***	**%**	Expert	7	2.2
< 20 years	47	15.0	No response	4	1.3
21–30 years	82	26.1			
31–40 years	41	13.1			
41–50 years	92	29.3			
51–60 years	43	13.7			
> 60 years	6	1.9			
No response	3	1.0			

Data were collected using a structured online questionnaire comprising 29 closed-ended, multiple-choice, Likert-scale, and open-ended items addressing different aspects of users' interactions with generative AI. Participants provided informed consent electronically by voluntarily proceeding to and completing the anonymous online questionnaire. Before participation, they were informed about the study's purpose, the voluntary nature of participation, and the scientific use of aggregated data. No written or verbal consent was collected.

While the questionnaire covered a broader range of issues related to AI use, the present analysis focused on experience with AI hallucinations (Q16–20), perceptions of AI transparency and explainability (Q9–15), verification and reliance strategies (Q23), and self-reported knowledge of AI limitations and risks (Q26). Open-ended questions allowed respondents to describe their experiences with AI hallucinations and the challenges they associated with AI-generated information.

ANALYSIS

The data were analysed in several stages using descriptive and inferential statistics. Given the predominantly ordinal and categorical nature of the survey data, non-parametric tests were used to examine relationships between the main variables. Descriptive statistics were first used to examine respondents' awareness and experiences of AI hallucinations, their perceptions of the severity of the problem, and the strategies they employed to mitigate the risk of inaccurate or fabricated AI-generated information. Frequencies and percentages were calculated for categorical and multiple-response items, while means and standard deviations were used for composite and Likert-scale measures where appropriate. Responses to Q9–Q15 were used to construct the Perceived AI Opacity Index, covering respondents' perceptions of AI transparency and explainability. The seven-item measure demonstrated good internal consistency (Cronbach's $\alpha = .84$), supporting its use as a composite index. Higher scores represented greater perceived opacity and lower perceived explainability. Relationships between the main variables were examined using Spearman's rank-order correlation coefficient (ρ), selected because of the ordinal character of several measures. The analysis focused on the relationships between exposure to AI hallucinations, perceived severity, AI opacity, self-reported knowledge of AI limitations, and verification behaviour. Chi-square (χ^2) tests of independence were used for categorical variables, including the relationship between familiarity with the concept of AI hallucinations and reported experience of the phenomenon. Cramér's V was used to assess the strength of these associations. Verification practices were examined separately using the multiple-response item Q23. External verification, control prompting, auxiliary use of AI, and absence of mitigation were coded as binary variables, allowing these strategies to be examined both descriptively and in relation to previous exposure to AI hallucinations. Finally, binary logistic regression was used to examine factors associated with external verification. External verification was entered as the dependent variable, with hallucination exposure, perceived severity, perceived AI opacity, self-reported knowledge of AI limitations, experience with generative AI, and frequency of AI use included as predictors. Continuous and ordinal predictors were standardised before estimation, allowing the odds ratios to represent the effect of a one-standard-deviation increase in each predictor.

RESULTS

Experiences and Awareness of AI Hallucinations

Encounters with inaccurate or fabricated AI-generated information were common among respondents. Of the 310 participants who provided a valid response, 156 (50.3%) reported encountering such content sporadically and a further 99 (31.9%) often. Overall, 82.3% had experienced inaccurate or fabricated AI-generated information at least occasionally. Direct experience was considerably more widespread than familiarity with the concept itself (Fig. 2). Only 132 respondents (42.6%) indicated that they had previously encountered the term AI hallucination. Conceptual awareness and reported experience were nevertheless significantly related, $\chi^2(3) = 42.52$, $p < .001$, with a moderate association (Cramér's $V = .37$). Respondents familiar with the term were more likely to report having encountered hallucinations, while

uncertainty about whether such encounters had occurred was more common among those unfamiliar with the concept.

Exposure to hallucinations was positively correlated with perceived severity, ρ = .44, p < .001. Respondents who encountered inaccurate or fabricated information more frequently were therefore more likely to regard hallucinations as a substantial problem.

Perceived AI Opacity and Epistemic Risk

Respondents' perceptions of AI transparency and explainability were measured using the seven-item Perceived AI Opacity Index (Q9–Q15), which demonstrated good internal consistency (Cronbach's α = .84). Higher scores represented greater perceived opacity and lower perceived explainability. The mean score was 2.96 (SD = 0.85), suggesting a moderate level of uncertainty regarding the transparency and explainability of generative AI systems. Greater exposure to inaccurate or fabricated AI-generated information was associated with higher perceived opacity, ρ = .22, *p* < .001. Although relatively modest, the relationship indicates that respondents who encountered hallucinations more frequently tended to perceive AI systems as less transparent and explainable. A stronger association was found between perceived opacity and the perceived severity of hallucinations, ρ = .42, *p* < .001. Exposure to hallucinations was also positively associated with self-reported knowledge of the limitations and risks of generative AI, ρ = .25, *p* < .001. Overall, users with greater exposure to hallucinations tended to report both greater awareness of AI limitations and stronger concerns about the reliability and transparency of generative AI.

Verification and Reliance Strategies

Most respondents reported taking active steps to reduce the risk of relying on inaccurate or fabricated AI-generated information. Among those who provided a valid response to Q23, 236 (77.1%) indicated that they verified AI-generated information using external sources, making this the most frequently reported strategy (Fig. 3.). A further 197 respondents (64.4%) reported using AI primarily in an auxiliary or supporting role, while 81 (26.5%) used control prompts intended to reduce the likelihood of hallucinated outputs. Only 20 respondents (6.5%) reported taking no specific action. As respondents could select more than one option, these categories were not mutually exclusive.

External verification was related to previous experience of AI hallucinations. Greater exposure to inaccurate or fabricated AI-generated information was positively associated with external verification, ρ = .28, *p* < .001. The relationship remained significant when respondents who answered “difficult to say” to the hallucination-exposure question were excluded, ρ = .21, *p* < .001, indicating that the association was not dependent on treating uncertain responses as an intermediate category. The reported strategies represent different responses to AI unreliability. External verification involves checking generated information against independent sources, while auxiliary use limits the extent to which AI serves as a primary source of information. Control prompting, in turn, attempts to reduce inaccurate outputs through interaction with the system itself. Overall, most respondents did not rely on AI-generated information without some form of additional scrutiny or limitation.

Predictors of External Verification

A binary logistic regression was conducted to determine whether the relationship between hallucination exposure and external verification remained when other relevant factors were considered simultaneously. The model included 306 complete cases, with hallucination exposure, perceived severity, perceived AI opacity, self-reported knowledge of AI limitations, experience with generative AI, and frequency of AI use included as predictors. Hallucination exposure was the only statistically significant independent predictor of external verification. Greater exposure was associated with higher odds of checking AI-generated information against external sources (OR = 1.86, 95% CI [1.34, 2.58], $p < .001$). A one-standard-deviation increase in hallucination exposure was therefore associated with approximately 86% higher odds of external verification after the remaining variables were taken into account.

Table 3. Binary logistic regression predicting external verification

Predictor	B	OR	95% CI for OR	*p*
Hallucination exposure	0.619	1.86	1.34–2.58	< .001
Perceived problem severity	0.164	1.18	0.84–1.66	.348
Perceived AI opacity	–0.038	0.96	0.70–1.33	.818
Knowledge of AI limitations	0.037	1.04	0.75–1.43	.822
AI experience	0.020	1.02	0.70–1.48	.914
AI usage frequency	0.096	1.10	0.78–1.56	.591

Note. N = 306 complete cases. Outcome variable: external verification (0 = no, 1 = yes).
Predictors were standardised before analysis; odds ratios represent a one-standard-deviation increase.
OR = odds ratio; CI = confidence interval. McFadden's pseudo-R^2 = .089.

As shown in Table 3, none of the remaining predictors independently predicted external verification. Perceived severity, perceived AI opacity, self-reported knowledge of AI limitations, general experience with generative AI, and frequency of AI use were not statistically significant. When considered simultaneously, direct experience of inaccurate or fabricated AI outputs was therefore more closely associated with verification behaviour than general experience with AI or frequency of use. Given the cross-sectional nature of the study, this result should not be interpreted as evidence that exposure to hallucinations causes users to adopt verification practices. It does, however, demonstrate an independent association between experience of AI unreliability and external verification behaviour.

DISCUSSION

The findings reveal a clear discrepancy between users' experience of AI hallucinations and their conceptual awareness of the phenomenon. While 82.3% of respondents reported encountering inaccurate or fabricated AI-generated information at least occasionally, only 42.6% were familiar with the term AI hallucination. Users may therefore recognise that an AI-generated response is problematic without understanding such failures as part of a broader limitation of generative AI. Conceptual awareness and reported experience were nevertheless significantly related, suggesting that familiarity with the phenomenon may help users recognise it when it occurs. This distinction is important for understanding AI literacy. Existing frameworks generally emphasise technical knowledge, critical evaluation, and understanding the capabilities and limitations of AI systems (Long & Magerko, 2020; Ng et al., 2021). The present findings support these competencies but also distinguish between experiencing AI failure and being able to conceptualise it. Recognising that a particular answer is inaccurate does not necessarily entail understanding why such errors occur, how frequently they may occur, or what they reveal about the limitations of generative AI. Practical experience alone should therefore not be treated as equivalent to critical understanding. The positive relationship between hallucination exposure and perceived severity ($\rho = .44$, $p < .001$) adds another dimension to this pattern. Respondents who encountered hallucinations more frequently were also more likely to regard them as a substantial problem. This connects perceptions of AI unreliability to direct experience rather than only to general knowledge or public discussion of AI risks. However, the cross-sectional design cannot establish whether repeated exposure leads users to perceive hallucinations as more serious or whether users who are already more attentive to AI limitations are more likely to identify such failures. The findings nevertheless show that experience, recognition, and conceptual understanding do not necessarily develop at the same rate. From the perspective of Critical AI Literacy, this gap is particularly relevant. Effective engagement with generative AI requires more than noticing that a system occasionally produces incorrect information; users must also understand that such failures are connected to the limitations of generative models and adjust their expectations accordingly. Critical AI Literacy therefore involves moving from the recognition of individual errors towards a broader understanding of the conditions under which AI-generated information can and cannot be treated as reliable. This transition cannot be assumed simply because users have accumulated experience with generative AI. The findings concerning perceived AI opacity further illustrate the difficulty users face when evaluating AI-generated information. Greater exposure to hallucinations was associated with higher perceived opacity ($\rho = .22$, $p < .001$), although the relationship was relatively modest. A stronger association was found between opacity and the perceived severity of hallucinations ($\rho = .42$, $p < .001$). Users who regarded generative AI as less transparent and explainable were therefore more likely to perceive hallucinations as a serious problem. This suggests that concerns about AI unreliability are connected not only to individual factual errors, but also to uncertainty about how AI-generated information is produced and how its credibility can be assessed. Generative AI presents information in a form that closely resembles knowledgeable human communication, and fluent language, coherent explanations, and confident responses may function as indicators of credibility even when the underlying information is inaccurate. Previous research shows that individuals frequently rely on fluency and ease of processing when assessing reliability (Tversky & Kahneman, 1974; Kahneman, 2011), while the illusory truth effect demonstrates that

perceived credibility does not necessarily correspond to factual accuracy (Fazio et al., 2015). The same linguistic characteristics that make generative AI useful and persuasive can therefore make its failures more difficult to recognise.

These findings can also be considered in relation to epistemic authority. Individuals necessarily depend on external sources of knowledge because they cannot independently verify every claim, traditionally relying on indicators such as expertise, institutional affiliation, authorship, or access to evidence (Hardwig, 1985; Goldman, 1999). Generative AI complicates this relationship because it can reproduce the linguistic appearance of expertise without possessing the epistemic agency normally associated with an expert. Palese's (2026) description of AI as an emerging epistemic infrastructure is particularly relevant: as these systems become embedded in everyday information practices, they may increasingly influence what users encounter and regard as credible, even though the basis of individual outputs is often difficult to establish. This does not mean that users regard AI as an unquestioned authority. Rather, the results point to a more complicated relationship in which AI remains useful while its credibility becomes conditional. Perceived opacity makes conventional judgments of authority more difficult because users cannot always determine why a particular answer was generated or what evidence supports it. Under such conditions, confidence in the presentation of information becomes an unreliable substitute for confidence in the information itself. The challenge is therefore not simply to recognise that AI can be wrong, but to determine when apparently authoritative information should be accepted and when it requires independent verification. This links perceived opacity to the verification practices observed in the present study.

The verification findings provide further insight into how users respond to the uncertainty associated with generative AI. External verification was the most commonly reported strategy, used by 77.1% of respondents, while 64.4% reported using AI primarily in an auxiliary role and 26.5% employed control prompts intended to reduce the likelihood of hallucinated outputs. These results suggest that awareness of AI unreliability does not necessarily lead users to reject the technology. Instead, many continue to use generative AI while introducing additional forms of scrutiny or limiting the role it plays in their information practices. This pattern is consistent with the concept of trust calibration, which emphasises the adjustment of reliance to the actual capabilities and limitations of a system rather than the maximisation of trust itself (Okamura & Yamada, 2020). Liebherr et al. (2026) similarly describe calibrated trust as something that develops through continued interaction, while Duan and Zhao (2024) distinguish between initial, situational, and learned trust. External verification, auxiliary AI use, and control prompting represent different responses to uncertainty, but all allow users to continue benefiting from AI without necessarily accepting its outputs at face value. Verification can therefore be understood not simply as fact-checking, but as one of the mechanisms through which reliance on AI is negotiated. Greater exposure to inaccurate or fabricated outputs was associated with more frequent external verification ($\rho = .28$, $p < .001$), and the relationship remained significant in the sensitivity analysis. More importantly, hallucination exposure was the only statistically significant independent predictor of external verification in the logistic regression model (OR = 1.86, 95% CI [1.34, 2.58], $p < .001$). This distinction is significant because experience of AI is not equivalent to experience of AI failure. Simply

using generative AI frequently, or considering oneself an experienced user, was not independently associated with external verification in the regression model. This suggests that interaction with generative AI does not automatically result in more critical patterns of use. Users may become increasingly proficient in operating AI systems without necessarily becoming more cautious about the information those systems provide. Direct encounters with their limitations appear to be more closely associated with verification behaviour than familiarity with the technology itself.

This finding adds an important dimension to existing discussions of AI literacy. He et al. (2023) show that appropriate reliance on AI depends partly on users' ability to assess both their own knowledge and the capabilities of the system. The present results suggest that such calibration may also be connected to encounters with system failure, which provide users with concrete evidence that fluent and confident outputs cannot always be treated as reliable. Verification should therefore be understood as part of a broader adjustment in the relationship between users and AI-generated information. Checking external sources introduces an independent point of reference; auxiliary use limits the authority granted to AI; and control prompting represents an attempt to manage uncertainty within the interaction itself. These practices do not eliminate the epistemic risks associated with generative AI, nor do they demonstrate that users have achieved perfectly calibrated trust. They do, however, show that reliance can become conditional: AI remains useful, but its outputs are no longer necessarily treated as sufficient evidence on their own. This movement from relatively unqualified reliance towards more selective and context-dependent use provides an empirical basis for the concept of Critical AI Literacy as epistemic adaptation developed in this study.

Taken together, the findings provide support for understanding Critical AI Literacy as a form of epistemic adaptation to generative AI. Existing approaches to AI literacy emphasise the importance of understanding AI systems, critically evaluating their outputs, and using them responsibly (Long & Magerko, 2020; Ng et al., 2021). The present study builds on these approaches by placing greater emphasis on how users respond to a particular characteristic of generative AI: its ability to produce information that appears credible while remaining potentially inaccurate. Effective AI use therefore requires users to recognise uncertainty and adjust the degree of authority they grant to AI-generated information. The notion of epistemic adaptation captures this process without implying that users simply become more distrustful of AI. Critical AI Literacy can therefore be understood as the capacity to use AI while maintaining an appropriate degree of epistemic distance from its outputs. This involves recognising that fluency does not guarantee accuracy, seeking additional evidence when necessary, and adjusting reliance according to the context in which AI is being used. At the same time, Critical AI Literacy should not transfer responsibility for managing hallucinations entirely to individual users. Respondents may develop strategies to protect themselves against unreliable outputs, but this does not remove the responsibility of developers, educational institutions, and other organisations involved in the design and deployment of generative AI. Users cannot reasonably be expected to compensate fully for systems whose decision-making processes and sources of information remain difficult to evaluate. The development of Critical AI Literacy should therefore complement, rather than replace, efforts to improve system transparency, responsible design, and appropriate governance (Mittelstadt et al., 2016; Floridi et

al., 2018). This has direct implications for education. Current discussions of AI literacy increasingly recognise that students need more than the technical ability to use generative tools effectively (Ismaili, 2024; Danyaro et al., 2025). The present findings reinforce this argument but also suggest that exposure to the limitations of AI may itself have educational value. Teaching students only how to generate better outputs risks strengthening operational competence without necessarily improving their ability to judge whether those outputs should be trusted. AI literacy education should therefore include opportunities to identify hallucinations, compare AI-generated claims with external evidence, examine how confident language can conceal uncertainty, and consider when AI should serve as a primary, auxiliary, or inappropriate source of information. The objective is not to teach students to distrust generative AI, but to develop the ability to determine when reliance is justified.

CONCLUSIONS

The findings of this research support the conceptualisation of Critical AI Literacy as a form of epistemic adaptation. Such literacy extends beyond technical competence and involves recognising the limits of AI-generated knowledge, questioning apparently authoritative outputs when their reliability cannot be established, and seeking independent evidence where necessary. This does not imply rejecting generative AI. Rather, it involves developing a more conditional relationship in which users can benefit from these systems without automatically granting epistemic authority to the information they provide. At the same time, the study is subject to several limitations. The convenience sample should not be treated as representative of the wider population of generative AI users, while the cross-sectional design prevents causal conclusions about the relationship between hallucination exposure and verification. Critical AI Literacy was also interpreted theoretically rather than measured as a separate psychometric construct. Future research could examine these relationships using longitudinal or experimental designs, more diverse samples, and validated measures of Critical AI Literacy.

Nevertheless, the findings have implications for how AI literacy is understood and taught. If experience of AI failure is more closely associated with verification than general experience with the technology, teaching users simply how to use generative AI effectively may not be sufficient. AI literacy should also develop the ability to recognise unreliable outputs, distinguish confidence from accuracy, and determine when independent verification is necessary. As generative AI becomes increasingly rooted in education, professional practice, and everyday information seeking, knowing how to use these systems will remain important. Equally important, however, will be knowing when their outputs should not be taken at face value.

Declarations

Competing Interests

The author(s) declare that there are no competing interests.

Funding Statement

This research received no external funding.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request. The data are not publicly available due to privacy and ethical considerations.

Declaration of Generative AI Use

The authors used ChatGPT (OpenAI) solely to check the manuscript for grammar and spelling. All suggested changes were reviewed and approved by the authors, who take full responsibility for the final content.

Ethics statement

The study consisted of an anonymous, minimal-risk online survey of adult participants. It involved no intervention, experimental manipulation, collection of personal identifying information, or sensitive personal data. Participation was voluntary, and respondents were informed that their anonymised responses would be analysed only in aggregate form for scientific purposes, including research publications.

Formal ethics approval was not sought because, at the time of data collection, this type of anonymous, non-invasive survey research was not subject to a prior institutional ethics-review requirement at WSB University. Participants provided informed consent electronically by choosing to proceed with the questionnaire. The research was conducted in accordance with the principles of the Declaration of Helsinki.

Author Contribution

Both authors contributted equally to the production of the manuscript.

References

1. Banerjee S, Agarwal A, Singla S (2025), August Llms will always hallucinate, and we need to live with this. In *Intelligent systems conference* (pp. 624–648). Cham: Springer Nature Switzerland
2. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623
3. Brown TB, Mann B, Ryder N et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst, 33
4. Danyaro K, Abdullahi S, Abdallah A, Chiroma H (2025) Hallucinations in Large Language Models for Education: Challenges and Mitigation. Int J Teach Learn Educ. *4*

5. Duan H, Zhao W (2024) The Effects of Educational Artificial Intelligence-Powered Applications on Teachers' Perceived Autonomy, Professional Development for Online Teaching, and Digital Burnout. Int Rev Res Open Distrib Learn 25:57–76
6. Dunning D (2011) The Dunning–Kruger effect: On being ignorant of one's own ignorance. Advances in experimental social psychology, vol 44. Academic, pp 247–296
7. Fazio L, Brashier N, Payne B, Marsh E (2015) Knowledge Does Not Protect Against Illusory Truth. J Exp Psychol Gen, *144*
8. Ferreira FA (2026) Systematic Review of Mitigating Epistemic Risks of LLMs as Educational Tutors: hallucinations and formative effects. Conference papers
9. Floridi L, Cowls J, Beltrametti M et al (2018) AI4People—An ethical framework for a good AI society. Mind Mach 28(4):689–707
10. Giddens A (1990) The Consequences of Modernity. Stanford University Press
11. Goldman AI (1986) *Epistemology and cognition* (Vol. 19, No. 86, p. 131). Cambridge, MA: Harvard University Press
12. Goldman AI (1999) Knowledge in a social world. Oxford University Press
13. Hardwig J (1985) Epistemic Dependence. The Journal of Philosophy. 82(7)
14. He G, Kuiper L, Gadiraju U (2023) Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. CHI '23: CHI Conference on Human Factors in Computing Systems. 1–18
15. Ismaili Y (2024) AI Literacy and Adaptive Learning in Moroccan Education: Advancing Critical Thinking and Personalized Learning. Research on Education and Media
16. Ji Z, Lee N, Frieske R et al (2023) Survey of hallucination in natural language generation. ACM-CSUR 55(12):1–38
17. Kahneman D (2011) Thinking, fast and slow. Farrar, Straus and Giroux
18. Kalai AT, Nachum O, Vempala SS, Zhang E (2026) Evaluating large language models for accuracy incentivizes hallucinations. Nature, 1–3
19. Kasneci E, Sessler K, Küchemann S et al (2023) ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individual Differences 103:102274
20. Li Z, Yi W, Chen J (2026) Accuracy paradox: Addressing epistemic, manipulative, and societal risks of hallucination in AI governance. Comput Law Secur Rev 61:106311
21. Liebherr M, Enkel E, Law ELC et al (2026) Dynamic calibration of trust and trustworthiness in AI-enabled systems. Int J Softw Tools Technol Transf 28:105–121
22. Liu C, Lin C (2026) Are AI hallucinations the same as error information? A qualitative exploration of hallucination tolerance and resistance strategies. Online Inf Rev 50(2):390–408
23. Long D, Magerko B (2020) What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16
24. Luhmann N (1979) Trust and Power. Wiley, New York

25. Luhmann N (1988) Tautology and paradox in the self-descriptions of modern society. Sociol Theory 6(1):21–37. https://doi.org/10.2307/201911
26. Marková I (2025) Epistemic trust and authority. Cultures Sci 8(2):133–139. https://doi.org/10.1177/20966083251324673
27. Maynez J, Narayan S, Bohnet B, McDonald R (2020) On faithfulness and factuality in abstractive summarization. *Proceedings of ACL 2020*, 1906–1919
28. Messeri L, Crockett M (2024) Artificial intelligence and illusions of understanding in scientific research. Nature 627:49–58. https://doi.org/10.1038/s41586-024-07146-0
29. Metzger MJ, Flanagin AJ (2015) Psychological approaches to credibility assessment online. In: Sundar SS (ed) The Handbook of the Psychology of Communication Technology. Wiley, Hoboken, NJ, pp 445–466
30. Mittelstadt BD, Allo P, Taddeo M et al (2016) The ethics of algorithms: Mapping the debate. Big Data Soc 3(2):1–21
31. Monteith S, Glenn T, Geddes JR, Peter C, Whybrow PC, Achtyes E, Bauer M (2024) Artificial intelligence and increasing misinformation. Br J Psychiatry 224(2):33–35
32. Ng DTK, Leung JKL, Chu SKW, Qiao MS (2021) AI literacy: Definition, teaching, evaluation and ethical issues. Computers Education: Artif Intell 58:504–509
33. Nowotny H (2024) The Illusion of Control: Living with Digital Others. *Global Perspectives, 5(1)*
34. Obiefuna P (2025) Hallucination and the Collapse of Epistemic Trust (September 14)
35. Okamura K, Yamada S (2020) Adaptive trust calibration for human-AI collaboration. PLoS ONE
36. Ooi KB, Tan GWH, Al-Emran M et al (2023) The potential of generative artificial intelligence across disciplines: perspectives and future directions. J Comput Inform Syst. Epub ahead of print 5 October.
37. OpenAI (2023) GPT-4 technical report. *arXiv preprint arXiv:2303.08774*
38. Ou M, Ho SS (2024) Does knowledge make a difference? Understanding how the lay public and experts assess the credibility of information on novel foods. Public Underst Sci, *33*(2)
39. Ou M, Zheng H, Zeng Y, Hansen P (2024) Trust it or not: Understanding users' motivations and strategies for assessing the credibility of AI-generated information. New Media Soc 28:507–529
40. Palese R (2026) Artificial Truth: Algorithmic Power, Epistemic Authority, and the Crisis of Democratic Knowledge. *Societies, 16*
41. Popper K (1959) The logic of scientific discovery
42. Quattrociocchi W, Capraro V, Perc M (2025) Epistemological fault lines between human and artificial intelligence
43. Rozenblit L, Keil F (2002) The misunderstood limits of folk science: an illusion of explanatory depth. Cogn Sci 26(5):521–562
44. Shin D, Lim JS, Ahmad N et al (2024) Understanding user sensemaking in fairness and transparency in algorithms: algorithmic sensemaking in over-the-top platform. AI Soc 39(2):477–490

45. Sperber D, ClÉment F, Heintz C, Mascaro O, Mercier H, Origgi G, Wilson D (2010) Epistemic Vigilance. Mind Lang 25:359–393
46. Strauß S (2021) Deep Automation Bias: How to Tackle a Wicked Problem of AI? Big Data Cogn Comput 5:18
47. Sun Y, Sheng D, Zhou Z et al (2024) AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanit Soc Sci Commun 11:1278
48. Tanchuk NJ (2026) AI Personalized Learning and the Risk of Epistemic Consolidation. *Studies in Philosophy and Education*, 1–16
49. Tsao J (2025) Trajectories of AI policy in higher education: Interpretations, discourses, and enactments of students and teachers. Artificial Intelligence, Computers and Education, p 100496
50. Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. Science 185:1124–1130
51. Yao X, Lin L (2026) Reframing digital literacy: current paradigms, core challenges, and future research directions. *Humanities and Social Sciences Communications*
52. Zawacki-Richter O, Marín VI, Bond M, Gouverneur F (2019) Systematic review of research on artificial intelligence applications in higher education. Int J Educational Technol High Educ, 16(1)

Figures

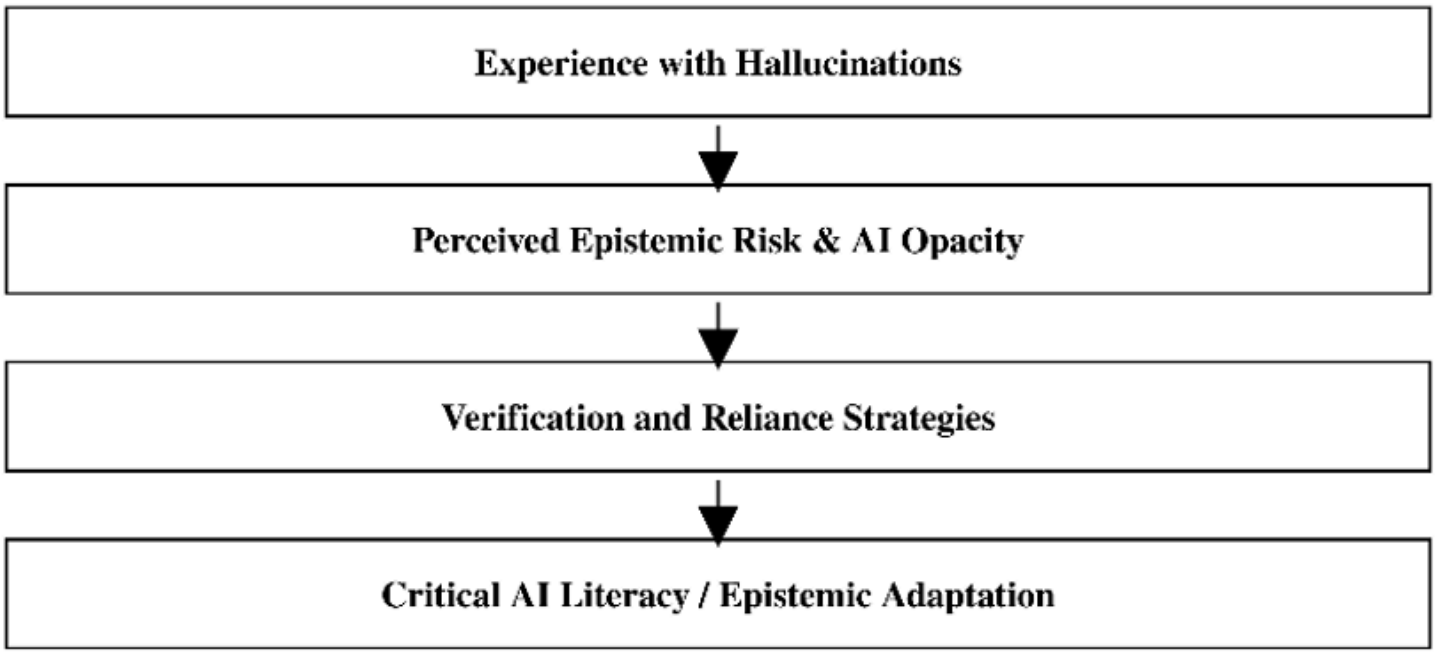


Figure 1

Conceptual model of Critical AI Literacy as epistemic adaptation to AI hallucinations

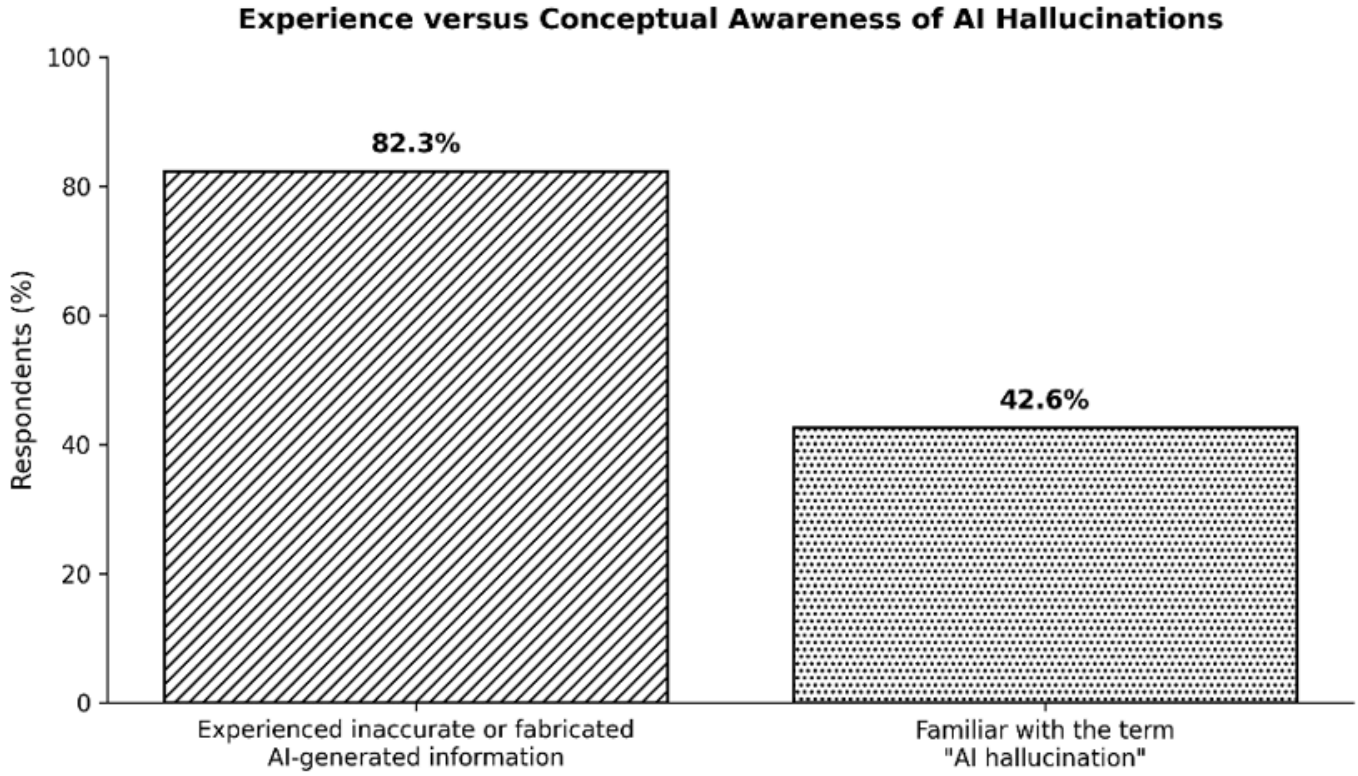


Figure 2

Experience versus Conceptual Awareness of AI Hallucinations

Exposure to hallucinations was positively correlated with perceived severity, $\rho = .44$, $p < .001$. Respondents who encountered inaccurate or fabricated information more frequently were therefore more likely to regard hallucinations as a substantial problem.

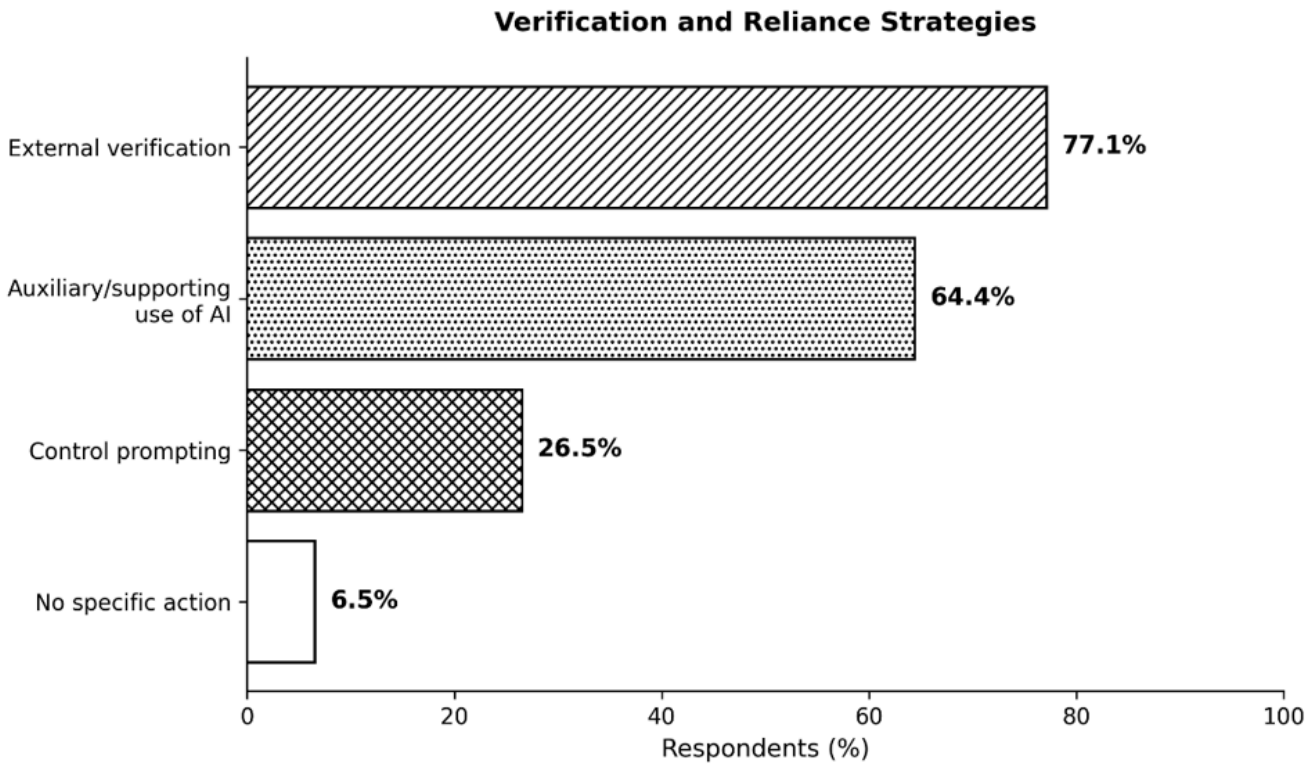


Figure 3

Verification and Reliance Strategies

Note. Respondents could select more than one strategy; percentages therefore do not sum to 100%.