IJEYD

Indonesian Journal of Education and Youth Development
e-ISSN: 3090-5680
Vol. 01, No. 03, 2025
Page: 110-123

OJS: https://ijeyd.org/index.php/ijeyd/index

Received: November 2, 2025	Accepted: November 20, 2025	Published: December 15, 2025

In Praise of AI Hallucinations: Re-imagining Critical Pedagogy and Truth Verification in the Age of Algorithmic Authority

Muhammad Afani Adam[1*], Hendra Novian[2]
[1,2]STAI Ma'arif Kalirejo Lampung Tengah, Indonesia
*🖂: afanieeadam@gmail.com

Abstrak

Integrasi cepat Model Bahasa Besar (LLM) dalam pendidikan tinggi telah memicu kekhawatiran luas tentang halusinasi, ketidakakuratan yang dihasilkan AI yang menantang integritas akademik. Namun, penelitian ini berpendapat bahwa ancaman epistemik yang lebih besar bukan terletak pada kesalahan AI, tetapi pada peningkatan kesempurnaannya, yang menumbuhkan otoritas algoritmik dan atrofi kognitif di antara siswa yang secara pasif mengonsumsi output yang akurat. Melalui metodologi penelitian perpustakaan yang sistematis, makalah ini mensintesis kerangka teoritis dari pedagogi kritis, pembelajaran transformatif, dan literasi informasi untuk mengusulkan paradigma kontra-intuitif: menginstrumentalisasi halusinasi AI sebagai aset pedagogis. Temuan menunjukkan bahwa sementara sistem akurasi tinggi menginduksi bias otomatisasi dan mengurangi kewaspadaan, output yang cacat dapat berfungsi sebagai dilema yang membingungkan yang mengaktifkan refleksi kritis dan kewaspadaan epistemik. Studi ini memperkenalkan pedagogi membaca forensik , mengadvokasi penggunaan strategis kesalahan AI untuk menumbuhkan keterampilan verifikasi dan otonomi analitis yang diperlukan untuk menavigasi ekosistem informasi yang dimediasi AI.

Kata kunci: AI generatif, pedagogi kritis, kewaspadaan epistemik, halusinasi AI, otoritas algoritmik, pembelajaran transformatif, literasi informasi.

Abstract

The rapid integration of Large Language Models (LLMs) in higher education has sparked widespread concern regarding hallucinations, AI-generated inaccuracies that challenge academic integrity. However, this study argues that the greater epistemic threat lies not in AI's errors, but in its increasing perfection, which fosters algorithmic authority and cognitive atrophy among students who passively consume accurate outputs. Through a systematic library research methodology, this paper synthesizes theoretical frameworks from critical pedagogy, transformative learning, and information literacy to propose a counter-intuitive paradigm: instrumentalizing AI hallucinations as pedagogical assets. Findings suggest that while high-accuracy systems induce automation bias and reduce vigilance, flawed outputs can function as disorienting dilemmas that activate critical reflection and epistemic vigilance. The study introduces a forensic reading pedagogy, advocating for the strategic use of AI errors to cultivate the verification skills and analytical autonomy necessary for navigating an AI-mediated information ecosystem.

Keywords: Generative AI, Critical Pedagogy, Epistemic Vigilance, AI Hallucinations, Algorithmic Authority, Transformative Learning, Information Literacy.

INTRODUCTION

The rapid integration of Large Language Models (LLMs) into academic environments has fundamentally transformed knowledge production, validation, and dissemination processes, positioning artificial intelligence as an epistemic infrastructure mediating teaching and learning practices across educational institutions worldwide (Chen, 2025). Contemporary discourse surrounding AI in higher education predominantly centers on mitigating technological failures, particularly addressing the phenomenon of hallucinations , instances where LLMs generate plausible yet factually incorrect or entirely fabricated information, which pose critical challenges to academic integrity and information literacy (Danyaro et al., 2024).

This prevailing narrative reflects what can be characterized as algorithmic authority, whereby users exhibit a tendency to trust algorithm outputs over human judgment due to the speed, coherence, and confident presentation of AI-generated responses, even when those responses contain fundamental errors. The epistemic danger inherent in increasingly accurate AI systems lies not merely in their potential for error, but paradoxically in their approaching perfection: as AI models achieve higher accuracy rates, users become progressively more passive, experiencing what recent scholarship identifies as cognitive atrophy, a progressive weakening of self-efficacy and analytical autonomy driven by habitual delegation of cognition to automated systems (Gupta, 2021; Kabashkin, 2025). This phenomenon manifests as reduced initiative, degraded error detection capabilities, and impaired takeover performance under conditions requiring critical evaluation, suggesting that perfect AI answers may paradoxically induce the very cognitive decline educators seek to prevent.

The theoretical foundation for understanding AI-human interaction in educational contexts draws upon multiple conceptual frameworks that illuminate the epistemic and pedagogical dimensions of this relationship. Epistemic Vigilance, defined as the ability and disposition to evaluate information critically before accepting it as knowledge, this serves as a cornerstone concept for analyzing how learners navigate AI-generated content (Sperber et al., 2010). This framework, originally developed in cognitive science and recently extended for educational applications, encompasses three evaluative dimensions: assessing the reliability of information sources, scrutinizing the validity of claims through rigorous scientific standards, and examining the receiver's cognitive biases and socioemotional influences. Transformative Learning Theory, Complementing the cognitive perspective, Mezirow's concept of the disorienting dilemma provides a pedagogical lens for understanding how encounters with problematic or contradictory information can catalyze critical reflection and perspective transformation (Code et al., 2022). The disorienting dilemma represents a psycho-cultural process wherein learners confront challenges to their existing frames of reference, precipitating critical self-examination and potentially fundamental shifts in meaning-making schemas. Then, Critical Pedagogy**, r**ooted in Freirean principles, this framework emphasizes problem-posing education wherein learners develop their power to perceive critically the way they exist in the world, striving toward conscientization through dialogue and critical interpretation of reality (Freire, 2021). These theoretical strands converge to suggest that pedagogical interventions designed to cultivate epistemic vigilance must deliberately incorporate moments of cognitive disruption that activate rather than suppress learners' critical faculties.

Despite growing scholarly attention to AI hallucinations in educational contexts, existing literature exhibits a significant gap in conceptualizing these technological imperfections as pedagogical assets rather than merely as problems requiring elimination. Current research predominantly focuses on technical solutions to mitigate or eliminate hallucinations to protect students from misinformation, emphasizing model refinement, prompt engineering optimization, and accuracy enhancement as primary objectives. This protective paradigm, while addressing legitimate concerns about academic integrity and information quality, fails to interrogate the pedagogical potential of strategically instrumentalizing AI errors as active components of critical thinking curricula.

Recent empirical work demonstrates that when pedagogically scaffolded, generative AI can transform students from passive users to critical evaluators, fostering strategies for bias detection and source validation (Rana et al., 2025). However, systematic investigations into how AI hallucinations might be deliberately repurposed as disorienting dilemmas that activate epistemic vigilance remain conspicuously absent from the literature. Moreover, while scholarship acknowledges the risks of cognitive atrophy associated with over-reliance on accurate AI systems, few studies have explored the counterintuitive proposition that human verification skills may be optimally developed through engagement with systems that lie convincingly rather than those that deliver truth with unwavering reliability. This gap represents a fundamental blind spot in our collective understanding of how flawed technology might paradoxically serve as superior pedagogical infrastructure for developing the very critical capacities necessary to navigate AI-mediated learning environments.

This study proposes a paradigm shift in conceptualizing AI hallucinations: repositioning them not as technological failures requiring elimination, but as pedagogically valuable disorienting dilemmas that can activate and cultivate critical thinking capacities in educational contexts. The primary objective is to demonstrate that human epistemic vigilance and verification skills are optimally sharpened through encounters with AI systems that generate convincing yet flawed outputs, rather than through reliance on systems that serve truth with minimal cognitive demand on the user.

Specifically, this research addresses three interconnected questions:

1. How does reliance on highly accurate AI systems affect student epistemic vigilance and critical evaluation capacities?
2. In what ways can AI hallucinations be strategically repurposed as pedagogical tools for teaching information literacy and critical thinking?
3. What theoretical frameworks support the use of flawed technology to enhance critical pedagogy and prevent cognitive atrophy?

By synthesizing scholarship across epistemic vigilance, transformative learning theory, critical pedagogy, and AI-human interaction, this study contributes to educational discourse by offering a conceptual framework for instrumentalizing technological imperfection as a catalyst for cognitive development. The anticipated contribution extends beyond theoretical innovation to practical pedagogy: by articulating how educators might deliberately integrate AI hallucinations into curricula as teachable moments rather than threats, this research provides a foundation for designing educational interventions that cultivate the analytical autonomy, metacognitive awareness, and critical literacy essential for navigating an increasingly AI-mediated information ecosystem.

METHOD

This library research adopts a systematic literature review methodology to synthesize existing theoretical frameworks, empirical findings, and pedagogical approaches relevant to understanding AI hallucinations as potential educational tools. The rationale for conducting literature-based rather than empirical research stems from the necessity to first establish a robust conceptual foundation by critically evaluating and integrating diverse scholarly perspectives across multiple disciplinary boundaries, including cognitive science, educational technology, critical pedagogy, and artificial intelligence ethics (Booth et al., 2021; Creswell & Creswell, 2017; Takona, 2024).

As emphasized in research methodology literature, comprehensive literature reviews serve essential functions: they situate new research within existing knowledge domains, identify gaps and inconsistencies in current understanding, establish theoretical or conceptual frameworks to ground subsequent investigations, and provide evidence and justification for novel research directions (Tavakol & O'Brien, 2023). Given the novelty of proposing AI hallucinations as pedagogical instruments, a paradigm shift that contradicts prevailing protective approaches, systematic synthesis of existing scholarship on epistemic vigilance, transformative learning, cognitive atrophy, and critical pedagogy becomes imperative to construct a theoretically coherent argument. Furthermore, library research enables the examination of cross-disciplinary insights that might otherwise remain siloed, facilitating the development of an integrative framework that bridges technical AI literature with educational theory and praxis.

Data collection utilizes major academic databases such as JSTOR, ERIC, Google Scholar, and the ACM Digital Library to aggregate relevant literature. Inclusion criteria prioritize peer-reviewed articles published from 2020 to the present for AI-specific discourse, while incorporating foundational texts for pedagogical theory, technical reports on LLM hallucination metrics, and philosophical works on digital authority. The subsequent analysis employs a thematic approach to identify recurring concepts regarding trust in technology, automation bias, and active learning, culminating in a synthesis that maps the mechanics of AI hallucination against the theoretical frameworks of Paulo Freire's Critical Pedagogy and Jack Mezirow's Transformative Learning Theory.

RESULTS AND DISCUSSION

The Psychology of Algorithmic Authority

Contemporary research reveals that human decision-makers exhibit systematic automation bias, defined as the propensity to favor suggestions from automated decision-making systems over contradictory information from non-automated sources, even when that information is accurate. This cognitive heuristic manifests as both errors of commission, accepting incorrect algorithmic recommendations, and errors of omission, failing to act without automated guidance. Empirical investigations across multiple domains demonstrate that automation bias is particularly pronounced when AI systems present high accuracy rates; paradoxically, as system reliability increases, users' critical vigilance correspondingly decreases. In mammography interpretation, for instance, radiologists across all experience levels demonstrated significant automation bias when presented with incorrect BI-RADS assessments from a purported AI system, with inexperienced readers' accuracy dropping from 79.7% to 19.8% when provided with erroneous AI suggestions (Gaube et al., 2021). The phenomenon operates through risk homeostasis mechanisms, wherein individuals calibrate their behavioral caution inversely to their perceived level of risk: when AI introduces a perceived level of accuracy or infallibility, decision-makers become more likely to accept suboptimal recommendations.

The conversational interface design of contemporary LLMs exacerbates this bias through what can be conceptualized as a user illusion of epistemic intimacy. The natural language interaction paradigm creates perceptual fluency that generates affective comfort, leading users to misinterpret confident linguistic presentation as indicative of factual reliability. Research on epistemic integrity in LLMs identifies epistemic miscalibration, the divergence between a model's linguistic assertiveness and its actual internal certainty, as a critical mechanism whereby high-confidence false statements mislead users on massive scales. Consequently, when AI systems achieve 99% accuracy, students effectively cease critical verification behaviors, creating what can be termed a passive consumption loop . This cognitive atrophy represents a more insidious threat to critical thinking development than overt inaccuracy: perfect AI answers induce learned helplessness in verification skills, whereas flawed systems necessitate continuous epistemic vigilance. Empirical evidence supports this counterintuitive relationship, moderate automation levels promote healthy trust calibration and improved decision-making accuracy, while high automation induces excessive reliance and diminished alertness.

Table 1. Comparison of Automation Effects

Dimension	**High-Accuracy AI Systems**	**Flawed/Moderate-Accuracy AI Systems**
User Vigilance	Significantly decreased; users cease verification behaviors	Enhanced; errors trigger critical evaluation

Cognitive Engagement	Passive consumption; cognitive atrophy	Active verification; skill development
Error Detection	Impaired; automation bias leads to acceptance of errors	Strengthened through necessity
Learning Outcomes	Reduced self-efficacy and analytical autonomy	Improved motivation, engagement, skill acquisition
Trust Calibration	Uncritical over-reliance	Healthy, appropriately calibrated trust

Re-framing Hallucinations: The Disorienting Dilemma

LLM hallucinations emerge from the fundamental architecture of these systems as stochastic parrots, entities that haphazardly stitch together sequences of linguistic forms according to probabilistic information about how they combine, without any reference to meaning. This characterization, introduced by linguist Emily Bender, captures the essence of probabilistic generation: LLMs operate through pattern matching in training data rather than factual retrieval from verified knowledge bases, producing outputs that are syntactically coherent and statistically plausible but semantically ungrounded (Bender et al., 2021). The stochastic component indicates determination by random, probabilistic distribution, meaning that even identical prompts may yield divergent outputs depending on sampling parameters. These systems are fundamentally prone to errors and biases, perpetuating stereotypes and problematic patterns embedded in training data while lacking transparency about inferential processes.

When reconceptualized through the lens of transformative learning theory, however, these hallucinations acquire profound pedagogical potential as catalysts for cognitive restructuring. Mezirow's framework posits that adult learning occurs through encountering disorienting dilemmas, situations that challenge existing perspectives and force critical self-examination of previously unquestioned assumptions (Mezirow, 2018). The theory describes transformation as occurring through a ten-stage process beginning with experiencing a disconcerting dilemma, proceeding through critical reflection on one's beliefs, and culminating in taking action based on newly integrated understandings. Transformation can be disruptive and uncomfortable precisely because learners are forced into seeing the world differently than previously accepted. Mezirow emphasizes that the goal of adult education is to facilitate autonomous thinking rather than to provide pre-packaged interpretations, learners must develop the capacity to make their own interpretations rather than uncritically accepting the beliefs and explanations of others.

Applied to AI pedagogy, this framework suggests that intentionally exposing students to plausible-sounding AI fabrications creates necessary cognitive crises that disrupt passive consumption patterns and activate epistemic vigilance. When students encounter AI-generated information that appears authoritative but contains subtle factual distortions, they experience the disorienting dilemma essential for transformative learning, their trust in algorithmic authority confronts contradictory evidence, precipitating critical reflection on how they evaluate information sources. Research in Swedish upper-secondary classrooms demonstrates this dynamic: students' ideas were influenced by biased AI-generated information presenting essentialist gender perspectives while marginalizing non-binary viewpoints, revealing AI's powerful agency in classroom discourse (Efimova & Nygren, 2025). This study identified two contrasting student orientations: AI optimism (uncritical acceptance) and AI vigilance (critical evaluation), with vigilance requiring both AI literacy to understand system limitations and clarity about task requirements to avoid superficial imitation. The pedagogical pivot lies in deliberately structuring encounters with AI hallucinations as opportunities for students to develop from optimistic consumers to vigilant evaluators, thereby transforming technological imperfection from liability into pedagogical asset.

A Pedagogy of Forensic Reading

Traditional information literacy frameworks such as the CRAAP test (Currency, Relevance, Authority, Accuracy, Purpose), developed in 2004 for evaluating information sources, demonstrate significant inadequacy when applied to generative AI outputs (Blakeslee, 2004). The CRAAP test's checklist approach encourages students to treat evaluation criteria as boxes to be ticked, rarely prompting them to leave the source under examination to gather contextual information. Critics note that students apply these criteria superficially without conducting the lateral verification necessary to assess source credibility in digital environments. When confronted with AI-generated content that exhibits surface-level currency (recent generation date), apparent relevance (tailored to prompts), pseudo-authority (confident presentation), fabricated accuracy (citation-like references), and ambiguous purpose (no transparent disclosure of generative processes), the CRAAP framework collapses.

In response to these limitations, information literacy scholarship increasingly advocates for lateral reading and adversarial reading approaches grounded in source investigation rather than content-focused evaluation. Lateral reading, operationalized through cognitive apprenticeship pedagogy, teaches learners to leave the text being evaluated, conduct parallel searches about the source's credibility, and trace claims to their origins before making credibility judgments. Experimental evidence with 312 participants demonstrates that lateral reading training based on cognitive apprenticeship, particularly when delivered through scalable written instructions, significantly enhances participants' abilities to identify misinformation (Fendt et al., 2023). The training emphasizes sourcing as a crucial strategy, addressing the tendency of users to overlook or superficially process source information, particularly on social networks where misinformation agents exploit identity obscuration.

Extending this approach to AI pedagogy, we propose a forensic reading framework that treats AI outputs as texts requiring adversarial scrutiny rather than passive consumption. This model incorporates what can be termed a reverse Turing test pedagogy, wherein students are assigned tasks requiring them to identify which portions of texts, arguments, or citations originated from AI generation versus human authorship (Valiaiev, 2024). In the reverse formulation, the human becomes the judge assessing whether content derives from machine or human origin, inverting Turing's original test where machines attempt to fool human evaluators. Pedagogical applications might include providing students with hybrid essays containing both authentic research and AI-generated fabrications, then requiring forensic analysis to detect hallucinated citations, logical inconsistencies, or stylistic anomalies indicative of probabilistic generation.

The critical pedagogical advancement lies in cultivating epistemic vigilance, not merely correcting AI fabrications, but understanding the generative mechanisms that produce them. This involves teaching students to interrogate why the AI fabricated specific information: Was it due to gaps in training data? Biases embedded in corpus composition? Probabilistic conflation of semantically proximate concepts? Inadequate constraint mechanisms in the model architecture? Extended epistemic vigilance frameworks encompass three evaluative dimensions applied to AI outputs: (1) assessing the reliability of the generative system as an information source, (2) scrutinizing the validity of claims through cross-referencing with authoritative sources, and (3) examining the user's own cognitive biases and socioemotional factors that might predispose acceptance of AI outputs (Bielik & Krell, 2025). By shifting educational focus from error correction to error etiology, forensic reading pedagogy transforms students from passive consumers who expect truth into active investigators who presume the necessity of verification.

Table 2. Comparative Frameworks for Information Literacy

Framework	Evaluation Focus	Student Activity	Effectiveness with AI	Key Limitation
CRAAP Test	Content-focused checklist (Currency, Relevance, Authority, Accuracy, Purpose)	Surface-level criteria assessment without lateral investigation	Inadequate, AI easily satisfies superficial criteria	Treats evaluation as checklist; fails to prompt external verification
Lateral Reading	Source credibility through parallel investigation	Leave text, search source reputation, trace claims to origins	Effective for misinformation detection with cognitive apprenticeship training	Less effective for evaluating truthful content; requires scaffolding
Forensic Reading (Proposed)	Adversarial scrutiny of generative mechanisms	Reverse Turing test exercises; identify AI vs. human content; analyze *why* hallucinations occur	Addresses AI-specific challenges; cultivates epistemic vigilance	Requires AI literacy and understanding of LLM architecture

The Role of Friction in Learning

Educational psychology scholarship on desirable difficulties, pioneered by Robert Bjork, establishes that learning conditions which introduce appropriate cognitive challenges, thereby slowing apparent performance gains, paradoxically support superior long-term retention and transfer compared to conditions that enable rapid, fluent performance (Bjork & Bjork, 2011).

Bjork distinguishes between perceptual fluencies that create comfortable feelings and measurable short-term performance gains (which students misinterpret as effective learning) versus effortful encoding and retrieval processes that seem uncomfortable but prove necessary for durable learning. Desirable difficulties include varying practice conditions to prevent contextual dependency, interleaving instruction rather than blocking by topic, spacing practice over time, and employing testing as a learning mechanism rather than merely assessment. Critically, Bjork emphasizes that not all difficulties are desirable, challenges must be surmountable with background knowledge and must trigger appropriate encoding and retrieval processes.

When applied to AI-mediated learning, this framework reveals that perfect AI systems remove essential cognitive friction, creating perceptual fluency that feels like learning while actually undermining the encoding and retrieval processes necessary for knowledge consolidation. AI systems that provide instant, accurate, comprehensive answers enable students to experience the satisfying feeling of information access without engaging the effortful cognitive processes that produce learning. Research on practice with reduced AI assistance demonstrates that partial automation during training leads to significantly better worker motivation, engagement, and skill acquisition compared to high automation conditions. The study's title, Practice With Less AI Makes Perfect , captures the counterintuitive finding that reducing automated support during skill development produces superior outcomes.

AI hallucinations, when pedagogically scaffolded, reintroduce desirable difficulties by requiring students to engage in verification processes that constitute the core competencies of information literacy and research methodology. The struggle to verify a complex, hallucinated citation, determining whether the referenced source exists, locating the actual publication if it does, assessing whether the citation accurately represents the source's argument, and understanding why the AI generated the fabrication, represents precisely the site where actual research skill develops. This verification labor involves: navigating scholarly databases, constructing effective search queries, evaluating source credibility, tracing citation chains, recognizing disciplinary conventions, and synthesizing information across multiple sources. Each of these constitutes a transferable skill that atrophies when students rely on AI systems that perform these cognitive operations automatically.

The pedagogical implication is that educators should deliberately calibrate AI accuracy levels to introduce optimal friction: systems that are sufficiently plausible to require serious engagement but sufficiently flawed to necessitate verification. This mirrors Bjork's insistence that desirable difficulties must be scaled to learners' competence, too little challenge produces no learning benefit, while excessive challenge overwhelms cognitive resources and becomes undesirable difficulty. In practical terms, this might involve: providing students with AI-generated literature reviews containing a known percentage of fabricated citations; requiring students to verify all claims before incorporating them into their work; and assessing students not on the accuracy of initial AI outputs but on the thoroughness and sophistication of their verification processes. Such pedagogical designs position AI hallucinations as deliberate provocations for epistemic engagement rather than technological failures to be eliminated, reframing imperfection as the necessary friction that transforms passive consumption into active learning.

CONCLUSION

The pursuit of frictionless accuracy in educational AI tools fundamentally misunderstands the cognitive mechanisms required for deep learning. As this study demonstrates, the elimination of error from algorithmic systems paradoxically eliminates the necessity for human critical engagement, fostering a dangerous dependency where students surrender their epistemic agency to the perfect machine. By reframing AI hallucinations not as technological defects to be patched but as disorienting dilemmas essential for transformative learning, we uncover a vital pedagogical opportunity. The cracks in the system, the fabrications, biases, and logical leaps, are precisely where critical inquiry enters; they force the student to shift from a passive consumer of information to an active forensic investigator of truth.

Consequently, the role of the educator in the age of algorithmic authority must shift from policing AI use to designing curricula that strategically incorporate technological imperfection. Rather than striving for tools that provide immediate, correct answers, educational practice should embrace a pedagogy of forensic reading that intentionally leverages AI errors to cultivate epistemic vigilance. In an era saturated with synthetic media, the mark of an educated mind is no longer the ability to retrieve information, but the capacity to detect fabrication and interrogate the provenance of claims. Ultimately, the flawed AI serves as the necessary gymnasium for the modern mind, providing the resistance required to build intellectual strength, whereas the perfect AI offers only the comfortable atrophy of the couch.

REFERENCES

Bender, E. M., 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. https://doi.org/10.1145/3442188.3445922

Bielik, T., & Krell, M. (2025). Developing and evaluating the extended epistemic vigilance framework. *Journal of Research in Science Teaching*, *62*(3), 869–895.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In *Psychology and the real world: Essays illustrating fundamental contributions to society* (Vol. 2, pp. 56–64).

Blakeslee, S. (2004). The CRAAP test. *Loex Quarterly*, *31*(3), 4.

Booth, A., Martyn-St James, M., Clowes, M., & Sutton, A. (2021). *Systematic approaches to a successful literature review*.

Chen, B. (2025). Beyond tools: Generative AI as epistemic infrastructure in education. *ArXiv Preprint ArXiv:2504.06928*.

Code, J., Ralph, R., & Forde, K. (2022). A disorienting dilemma: Teaching and learning in technology education during a time of crisis. *Canadian Journal of Science, Mathematics and Technology Education*, *22*(1), 170–189.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Danyaro, K. U., Abdullahi, S., Abdallah, A. S., & Chiroma, H. (2024). Hallucinations in Large Language Models for Education: Challenges and Mitigation. *International Journal of Teaching, Learning and Education*, *4*(6), 639993.

Efimova, E., & Nygren, T. (2025). Classroom Discussions of Social Issues in the Age of Generative AI: Epistemic Vigilance Against Bias and Bullshit. *The Journal of Social Studies Research*.

Fendt, M., Nistor, N., Scheibenzuber, C., & Artmann, B. (2023). Sourcing against misinformation: Effects of a scalable lateral reading training based on cognitive apprenticeship. *Computers in Human Behavior*, *146*, 107820.

Freire, P. (2021). *Education for critical consciousness*.

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., & Ghassemi, M. (2021). Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digital Medicine*, *4*(1), 31.

Gupta, S. (2021). *Keep sharp: Build a better brain at any age*. Simon & Schuster.

Kabashkin, I. (2025). Cognitive Atrophy Paradox of AI–Human Interaction: From Cognitive Growth and Atrophy to Balance. *Information*, *16*(11), 1009.

Mezirow, J. (2018). Transformative learning theory. In *Contemporary theories of learning* (pp. 114–128). Routledge.

Rana, V., Verhoeven, B., & Sharma, M. (2025). Generative AI in design thinking pedagogy: Enhancing creativity, critical thinking, and ethical reasoning in higher education. *Journal of University Teaching and Learning Practice*, *22*(4), 1–22.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393.

Takona, J. P. (2024). Research design: qualitative, quantitative, and mixed methods approaches / sixth edition. *Quality and Quantity*, *58*(1), 1011–1013. https://doi.org/10.1007/S11135-023-01798-2/METRICS

Tavakol, M., & O'Brien, D. (2023). The importance of crafting a good introduction to scholarly research: strategies for creating an effective and impactful opening statement. *International Journal of Medical Education*, *14*, 84.

Valiaiev, D. (2024). Detection of machine-generated text: Literature survey. *ArXiv Preprint ArXiv:2402.01642*.