International Educational Review

https://int-er.com

Algorithmic Authority Bias in AI-Assisted Learning: An Integrative Cognitive Framework

Sara D Sony

Northwest Missouri State University, USA

Koksal Mus

Worcester Polytechnic Institute, USA

Abstract: The increasing reliance on generative artificial intelligence (AI) in education is reshaping how learners' access, evaluate, and internalize knowledge. While AI tools enhance accessibility, they also introduce cognitive risks by encouraging learners to bypass critical thinking. This paper introduces Algorithmic Authority Bias (AAB), a conceptual framework that explains how learners assign unwarranted epistemic authority to AI-generated responses on the basis of their fluency, immediacy, and coherence rather than their accuracy or underlying expertise. Building on established theories of authority bias, automation bias, and fluency-based judgment, and integrating recent empirical work on metacognitive laziness, cognitive offloading, and the illusion of explanatory depth in AI-assisted learning, the paper foregrounds the role of cognitive offloading, whereby learners delegate cognitive tasks to AI systems, reducing their active engagement and metacognitive awareness. This over-reliance can produce "synthetic mastery" — a false sense of understanding in which learners accept plausible-sounding answers without engaging in conceptual reasoning. By developing a model that links specific AI features to cognitive mechanisms, the paper articulates a set of testable propositions describing how fluency, perceived authority, and offloading jointly contribute to the acceptance of responses that appear correct yet lack conceptual depth. It then considers implications for instructional design, AI system development, and assessment strategies, and outlines directions for empirical research that build on, rather than duplicate, recent experimental findings. Ultimately, the paper contributes to understanding how AI is reshaping epistemic processes in education and calls for a more critical approach to AI-assisted learning.

Keywords: Algorithmic Authority Bias; Artificial Intelligence; False Understanding; AI-Assisted Learning; Cognitive Bias; Educational Assessment.
DOI: https://doi.org/10.58693/ier.415

Introduction

The widespread adoption of artificial intelligence (AI) tools, particularly large language models (LLMs) such as ChatGPT, is significantly reshaping the educational landscape across diverse learning environments (Kasneci et al., 2023; Dwivedi et al., 2023; O'Sullivan et al., 2025; Nguyen et al., 2024). These tools are highly valued because they provide personalized, immediate, and fluent responses, making learning more accessible and efficient (Mogavi et al., 2024). However, their swift integration also raises critical concerns about students' ability to evaluate and assess the reliability of AI-generated content (Sperber et al., 2010; Efimova & Nygren, 2026). One concern is algorithmic bias, where AI systems may unintentionally reinforce existing inequities and shape learning outcomes (Baker & Hawn, 2022). As AI becomes ubiquitous in education, it is essential not only to examine its practical benefits but also to consider the cognitive and epistemic risks it introduces, particularly in how students engage with and internalize knowledge.

Emerging evidence suggests that students often over-rely on AI outputs, accepting them without adequate critical evaluation (Risko & Gilbert, 2016; Kasneci et al., 2023; Fan et al., 2024). This overreliance is largely driven by

cognitive offloading and the high fluency of generative systems, which can create the illusion of accuracy (Oppenheimer, 2008). Recent experimental work documents the consequences directly: students using ChatGPT showed improved performance scores but demonstrated no significant gains in knowledge transfer - a pattern labelled "metacognitive laziness" (Fan et al., 2024). Additional studies using Electroencephalography (EEG), a non-invasive method for measuring the brain's electrical activity, found that ChatGPT users displayed the lowest neural engagement across writing groups (Kosmyna et al., 2025), while AI-assisted research reduced cognitive load but produced lower-quality arguments (Stadler et al., 2024). These findings indicate that the cognitive risks of AI-assisted learning are not merely theoretical.

Traditional models such as authority bias (Cialdini & Goldstein, 2004) and automation bias (Parasuraman & Riley, 1997; Mosier & Skitka, 1996) provide partial explanations for these phenomena. Authority bias describes a tendency to defer to perceived expertise; automation bias describes overdependence on system recommendations. However, these models were developed prior to the rise of generative AI and do not fully account for the unique linguistic and conversational nature of modern AI systems. In particular, they neglect how *fluency* — the ease with which information is processed — can be mistaken for epistemic credibility, leading learners to equate clarity with understanding (Oppenheimer, 2008; Reber & Unkelbach, 2010).

Recent theoretical and empirical work has begun to address these limitations (i.e., the inadequacy of traditional authority and automation bias models in accounting for the unique fluency and conversational nature of generative AI). For instance, Jose et al. (2025) argue that generative AI functions as a "surrogate knower" that displaces traditional epistemic authority in classrooms. Additionally, Tanchuk et al. (2025) propose normative criteria for evaluating epistemic trustworthiness in AI tutors, and Pandey et al. (2025) developed a six-factor instrument for measuring epistemic trust in generative AI in higher education. Mehta et al. (2024) connect the *illusion of explanatory depth* (IOED; Rozenblit & Keil, 2002) directly to LLM use. Nevertheless, these contributions remain fragmented. No integrated cognitive model yet connects specific features of generative AI outputs —such as fluency, immediacy, structured coherence— to the cognitive mediators (authority heuristics, fluency-based judgment, cognitive offloading, weakened metacognition) that produce inflated epistemic trust and superficial learning.

To fill this integrative gap, the present paper introduces Algorithmic Authority Bias (AAB), a conceptual framework that explains how learners attribute unwarranted epistemic authority to AI-generated outputs based primarily on surface features such as fluency, coherence, and immediacy, rather than on actual epistemic validity. Rather than replacing existing constructs, AAB integrates and extends them to better account for the unique dynamics of conversational AI environments. Over time, these dynamics foster what this paper terms *false understanding* — a condition in which students generate correct or plausible answers without possessing the underlying reasoning required for genuine conceptual expertise (Chi, 2009; Balta, 2026). The framework provides a unified account of how generative AI reshapes epistemic judgment in education and offers a set of testable propositions to guide future empirical work.

Literature Review

Theoretical Foundations: Authority, Automation, and Cognitive Offloading

Human cognition is shaped by a tendency to defer to perceived authority, particularly under conditions of uncertainty. Individuals often rely on external sources, such as experts or institutions, to reduce cognitive effort (Milgram, 1963; Cialdini & Goldstein, 2004). In educational contexts, this deference frequently manifests as an acceptance of transmitted knowledge without deep evaluation, thereby limiting critical engagement and independent reasoning (Freire, 1970; Chi, 2009; Kirschner et al., 2006).

A closely related mechanism is *automation bias*. In this bias, individuals tend to over-rely on system outputs and reduce independent verification, especially when automated systems are perceived as efficient, objective, and reliable (Parasuraman & Riley, 1997; Mosier & Skitka, 1996; Lee & See, 2004; Parasuraman & Manzey, 2010). This tendency has been widely observed in decision-making contexts where users often follow algorithmic recommendations even when those recommendations conflict with their own judgment.

A third key strand is *fluency-based trust*, also known as a fluency heuristic. In this context, the ease of processing information serves as a powerful indicator of truth and credibility (Oppenheimer, 2008; Reber & Unkelbach, 2010; Hertwig et al., 2008). In AI-assisted learning environments, this fluency heuristic becomes particularly influential because generative AI responses are typically immediate, grammatically well-structured, and delivered with linguistic confidence. These qualities strongly reinforce perceptions of correctness, even in the absence of verification.

Finally, *cognitive offloading* refers to the delegation of cognitive tasks to external resources (Risko & Gilbert, 2016). Generative AI introduces a qualitatively new form of offloading. Learners can now delegate not only retrieval but also explanation, reasoning, and problem-solving processes to systems that simulate epistemic agency through natural language generation.

Taken together, these four foundations — authority bias, automation bias, fluency-based judgment, and cognitive offloading — offer complementary but individually incomplete explanations of learner behavior in AI contexts. None of them, in isolation, fully captures the specific interplay of cues, cognitive shortcuts, and learning consequences that arises when learners interact with conversational generative AI.

Generative AI and the Transformation of Epistemic Interaction in Education

LLMs have moved beyond simple search tools to become central to student learning experiences. These models support tasks such as assignment completion, concept clarification, and facilitating comprehension (Kasneci et al., 2023). Although these features improve efficiency and access, they also fundamentally alter the nature of epistemic engagement. That is, they change how students engage with knowledge. The rapid and fluent nature of AI-generated explanations can significantly reduce the cognitive effort students invest in problem-solving, justification, and self-

explanation. These steps are central to meaningful learning (Risko & Gilbert, 2016; O'Sullivan et al., 2025). Consequently, learners often transition from active knowledge construction to a reliance on AI-generated answers. This effectively positions the AI as a "surrogate knower" (Jose et al., 2025), displacing students' own role in actively constructing understanding.

Building on this shift from active learning to reliance on AI, another important concern is the surface qualities of AI-generated content. Even when AI-generated responses contain incomplete reasoning or factual inaccuracies, they are typically fluent, well-structured, and delivered with high confidence s (Ji et al., 2023). This combination of fluency and confidence can mislead students into overestimating the correctness and depth of explanations, creating an inflated sense of understanding (Efimova & Nygren, 2026). In conceptually demanding tasks, students may mistake well-presented explanations for genuine comprehension. Over time, repeated exposure to such interactions may shift learning from active cognitive engagement toward passive acceptance of AI-generated content.

Moreover, these concerns are reinforced by issues of bias and fairness in educational AI systems. For example, Idowu et al. (2024) demonstrate that algorithmic bias in student progress monitoring can worsen inequities in learning outcomes. Similarly, Kizilcec and Lee (2022) highlight that algorithmic fairness significantly influences students' learning experiences and trust in AI systems. As a result, generative AI is not a neutral tool but an active force that reshapes how learners interpret, trust, and internalize academic knowledge.

Recent Empirical and Theoretical Work on Epistemic Trust and Cognitive Engagement in AI-Assisted LearningText

In recent years, a large and rapidly growing body of research (2024–2026) has begun to document the cognitive and epistemic consequences of AI-assisted learning. These studies provide valuable empirical and conceptual support for the framework presented in this paper. In this section, we organize the existing literature into four strands. The first two strands focus on empirical evidence, while the last two strands address conceptual and measurement work.

Empirical evidence for cognitive offloading and reduced engagement

Fan et al. (2024) conducted experimental research with 117 university students who revised essays using different types of support, including ChatGPT, human expert feedback, checklist tools, and no support. Students who used ChatGPT produced better essays in the short term; however, they showed no real improvement in long-term knowledge retention or the ability to apply what they learned (Fan et al., 2024). The researchers introduced the idea of metacognitive laziness (Fan et al., 2024). This means students relied so heavily on AI that they stopped using important thinking processes such as planning monitor and evaluation (Fan et al., 2024).

Building on this finding, Kosmyna et al. (2025) measured brain activity using EEG during essay writing. Students who used ChatGPT showed the lowest level of neural engagement compared to other groups (Kosmyna et al., 2025).

These effects continued even when students later worked without AI (Kosmyna et al., 2025). Similarly, Stadler et al. (2024) found that ChatGPT support reduced mental effort but produced weaker arguments. Additionally, Lee et al. (2025) reported that cognitive offloading is connected to less critical thinking, especially among younger students. Together, these studies show that the risks of reduced engagement are real under AI usage.

Empirical work on the illusion of understanding

Recent studies have shown that students using LLMs often believe they understand topics more deeply than they actually do (Kumar et al., 2026; Belghith et al., 2024). This phenomenon is known as the illusion of explanatory depth (Rozenblit & Keil, 2002). Moreover, later experiments confirmed that AI-generated explanations can strengthen this illusion (Chromik et al., 2021). Students feel they have mastered the material when, in fact, their understanding remains shallow. This research provides strong support for the concepts of synthetic mastery and false understanding that are central to the AAB framework.

Conceptual work on epistemic authority in AI-mediated learning

As discussed in earlier sections, Jose et al. (2025) and Tanchuk et al. (2025) examine how generative AI functions as a surrogate knower and offer frameworks for assessing its trustworthiness. Other researchers (Jakesch et al., 2023) have shown that fluent, well-written AI text is often judged as more credible than human-written text even when the information is not more accurate.

Measurement work on epistemic trust

Pandey et al. (2025) created and tested a new survey tool called the Epistemic Trust in Generative AI in Higher Education (ETGAI-HE) scale. This tool measures student trust in AI according to factors such as reliability, predictability, and user control (Pandey et al., 2025). It offers a useful way to study the ideas presented in the AAB framework.

In summary, the studies reviewed in this section demonstrate that the cognitive and epistemic risks of AI-assisted learning are not merely theoretical. They are supported by empirical data and conceptual analysis; however, most of this research examines the issues in separate ways. What is still missing is an integrated model that clearly connects specific features of AI outputs, such as fluency, speed, and coherence, to the mental processes that lead to over-trust and shallow learning. The AAB framework introduced in this paper aims to fill this important gap.

Conceptual Limitations of Existing Frameworks

The foundations and recent work reviewed above each illuminate important parts of the picture. However, none of them alone captures the full pathway from AI features to learning outcomes.

Authority bias and automation bias explain why students defer to external sources (Cialdini & Goldstein, 2004; Parasuraman & Riley, 1997), yet they do not show how linguistic fluency drives that deference in conversational AI

systems. Fluency research clarifies why ease of processing increases credibility judgments (Oppenheimer, 2008; Reber & Unkelbach, 2010), but it remains disconnected from the specific conditions of conversational AI interaction. Cognitive offloading research describes the delegation of cognitive work (Risko & Gilbert, 2016) yetit does not explain how learners come to trust the resulting outputs as authoritative. IOED captures the self-misattribution of understanding (Rozenblit & Keil, 2002), but it does not address the specific AI input features that trigger this effect. Finally, philosophical and normative accounts of AI epistemic authority (Jose et al., 2025; Tanchuk et al., 2025) describe what is happening at the level of classroom epistemology, yet they do not provide a cognitive-mechanism model.

What is still missing is a single integrated construct. This construct should (a) identify the specific AI input features that cue epistemic authority, (b) specify the cognitive mediators these cues activate, (c) characterize the recursive interaction among these mediators, and (d) connect them to defined learning outcomes. The AAB framework is offered as that integrative construct.

Algorithmic Authority Bias (AAB): Conceptual Integration

This section introduces Algorithmic Authority Bias (AAB), a mediating cognitive construct, as the central conceptual contribution of this paper. It helps explain why learners often treat AI-generated outputs as authoritative when those outputs sound fluent, look well-organized, and arrive quickly, even when the learner has not verified the accuracy of the information. AAB is not a new or primitive bias. Instead, it brings together four established mechanisms from cognitive and educational psychology: authority heuristics, fluency-based judgment, cognitive offloading, and reduced metacognitive monitoring. The framework places these mechanisms within the specific context of conversational AI use in learning environments.

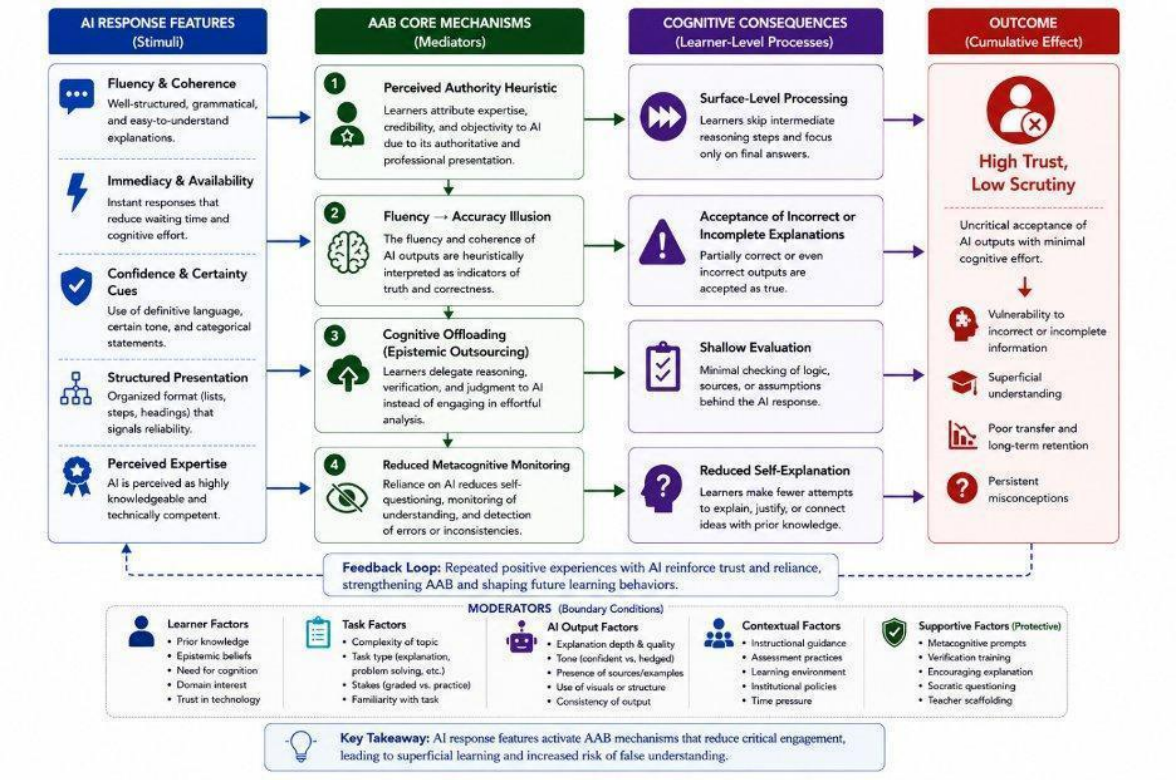
Figure 1 shows the full pathway through which AAB operates. The pathway begins on the left with AI response features, which act as stimuli. These features include fluency and coherence, immediacy and availability, confidence and certainty cues, structured presentation, and perceived expertise. These stimuli activate four core AAB mechanisms: the perceived authority heuristic, the fluency-to-accuracy illusion, cognitive offloading, and reduced metacognitive monitoring. The mechanisms then produce cognitive consequences for the learner. These consequences include surface-level processing, acceptance of incorrect or incomplete explanations, shallow evaluation, and reduced self-explanation. Together, these consequences result in the cumulative outcome of high trust paired with low scrutiny. A feedback loop captures the recursive reinforcement of this cycle. When students have repeated positive experiences with AI, their trust grows stronger and their reliance increases, which shapes future learning behaviors. Finally, moderators act as boundary conditions and fall into five categories: learner, task, AI output, contextual, and supportive factors. These moderators determine how strong AAB effects will be in any given learning environment, and they specify the conditions under which AAB effects strengthen or attenuate.

Although the mechanisms shown in Figure 1 share some components with other constructs already in the literature, AAB is conceptually distinct from each of them. Table 1 compares AAB with five adjacent constructs: authority bias, automation bias, fluency-based trust, IOED, and the "surrogate knower" concept introduced by Jose et al. (2025), which describes how AI displaces the teacher's traditional role as a source of knowledge in the classroom. The comparison uses four dimensions: the triggering cue, the primary mechanism, the cognitive scope, and the predicted outcome.

The contribution of AAB is therefore not novel at the level of any single mechanism. Instead, it lies in the explanation of how these mechanisms interact in conversational AI environments. Generative AI systems produce responses that are linguistically fluent, structurally coherent, and confidently articulated. These features act as surface-level cues of expertise. They lead learners to infer correctness from clarity. When fluency is misattributed to epistemic accuracy rather than to surface processing ease (Oppenheimer, 2008; Reber & Unkelbach, 2010), learners may accept AI-generated explanations as authoritative even when the underlying reasoning is incomplete or partially incorrect (Ji et al., 2023).

Figure 1

An Integrated Mechanistic Framework of Algorithmic Authority Bias (AAB) in AI-Assisted Learning.



The proximal cognitive outcome of AAB is the emergence of synthetic mastery (Balta, 2026). This means that learners experience a subjective sense of understanding without possessing fully developed conceptual structures (Balta, 2026).

Synthetic mastery is the pathway toward false understanding. False understanding is defined as the ability to generate correct or plausible answers without possessing the underlying reasoning needed for genuine conceptual expertise. The two constructs are distinct but related. Synthetic mastery refers to the in-the-moment subjective state of comprehension. False understanding refers to its durable behavioral expression, namely, performance without transfer. Both constructs are closely related to IOED in AI contexts (Mehta et al., 2024). However, they are framed here as outcomes of a specific mechanistic pathway rather than as a generic metacognitive illusion.

It is also important to acknowledge counterevidence. Dietvorst et al. (2015) documented algorithm aversion, in which users reject algorithmic outputs after observing errors. The AAB account is not incompatible with this finding. AAB predicts inflated trust under conditions of high fluency, immediacy, and coherence. However, AAB also allows that error salience, and prior failure can lower trust below baseline. In this way, AAB and algorithm aversion operate as opposing forces. Their balance depends on the quality of AI output, the visibility of errors, learner expertise, and task type.

Cognitive Mechanisms, Boundary Conditions, and Testable Propositions

AAB operates through four interacting cognitive mechanisms that shape how learners process and evaluate AI-generated information. These mechanisms have been studied individually in different bodies of research, but AAB describes how they work together in conversational AI environments. The four mechanisms are:

1. The perceived authority is heuristic. Learners attribute expertise to confident, well-structured outputs without verifying them (Cialdini & Goldstein, 2004).

2. The fluency-to-accuracy illusion. Learners interpret linguistically fluent and coherent explanations as more accurate than they are (Oppenheimer, 2008; Reber & Unkelbach, 2010).

3. Cognitive offloading. Learners delegate explanation, inference, and problem-solving to AI systems rather than engaging in independent reasoning (Risko & Gilbert, 2016; Gerlich, 2025).

4. Reduced metacognitive monitoring. Learners become less accurate in evaluating their own understanding when they are exposed to fluent and confident AI outputs (Flavell, 1979; Chi, 2009; Fan et al., 2024).

These four mechanisms reinforce one another in a recursive cycle. Fluency-based judgment increases epistemic trust in AI outputs. Greater trust encourages more cognitive offloading. Greater offloading further reduces opportunities for metacognitive reflection and error detection. In other words, each mechanism strengthens the conditions that activate the others. Over time, this recursive interaction contributes to an illusion of explanatory depth (Rozenblit & Keil, 2002) and produces the pathway from synthetic mastery to false understanding.

Table 1

AAB compared with adjacent constructs.

Construct	Triggering cue	Primary mechanism	Cognitive scope	Predicted outcome
Authority bias	Perceived expertise of source	Deference heuristic	Source evaluation	Acceptance of expert claims
Automation bias	Output of an automated system	Reduced verification	System reliance	Errors of omission / commission
Fluency-based trust	Ease of processing	Truth heuristic	Single judgment	Inflated credibility rating
IOED (in AI)	External availability of explanations	Self-misattribution of understanding	Self-assessment	Overestimation of own understanding
Surrogate knower (Jose et al., 2025)	AI-generated content in classroom	Displacement of dialogic justification	Classroom epistemology	Erosion of teacher authority
Algorithmic Authority Bias (AAB)	Fluency, immediacy, and structured output	Recursive interaction of the perceived authority heuristic, fluency-to-accuracy illusion, cognitive offloading, and reduced metacognitive monitoring	Both source and self-evaluation in conversational AI	Synthetic mastery leading to false understanding

Testable propositions. Based on this account, we derive six propositions to guide empirical research:

- P1. Higher fluency in AI output increases learners' epistemic trust independently of factual accuracy.
- P2. Immediacy and structured presentation of AI output amplify the effect of fluency on epistemic trust (an interaction effect).

- P3. AAB mediates the relationship between AI input features (fluency, immediacy, and structure) and the depth of learners' conceptual processing.
- P4. Greater AAB is associated with increased cognitive offloading and reduced metacognitive monitoring.
- P5. Synthetic mastery is a distinguishing behavioral signature of AAB. It can be operationalized as a pattern in which subjective understanding ratings match objective performance on near-transfer tasks but diverge on far-transfer tasks.
- P6. Prior domain knowledge, AI literacy, and epistemic disposition moderate susceptibility to AAB. Learners with higher knowledge and higher AI literacy are expected to show weaker AAB effects.

Boundary conditions. AAB is unlikely to apply uniformly across all learners, tasks, and contexts. As shown in Figure 1, five categories of moderators determine the strength of AAB effects in any given learning environment.

Learner factors shape individual susceptibility to AAB. Prior knowledge helps learners detect inconsistencies in AI explanations. Epistemic beliefs influence the willingness to question authoritative-looking outputs. Need for cognition predicts engagement in effortful evaluation. Domain interest motivates deeper processing. Finally, baseline trust in technology affects the default level of reliance on AI outputs. Together, these factors mean that high-knowledge, high-AI-literacy, and epistemically vigilant learners are likely to show weaker AAB effects (Sperber et al., 2010).

Task factors determine the cognitive demands placed on learners and, in turn, the value of cognitive offloading. Topic complexity, task type (for example, explanation versus problem-solving), stakes (for example, graded versus practice), and learner familiarity with the task all influence whether AI use support or undermines learning. Routine procedural tasks may benefit from offloading without producing false understanding. By contrast, conceptually demanding and unfamiliar tasks are the most vulnerable to AAB effects.

AI output factors moderate the strength of the cues that trigger AAB. Explanation depth and quality, tone (confident versus hedged), the presence of sources or examples, the use of visuals or structure, and the consistency of outputs across responses all shape learners' perceptions of authority and accuracy. When AI errors are visible and outputs hedge appropriately, algorithm aversion may dominate over AAB (Dietvorst et al., 2015; Kim et al., 2024).

Contextual factors situate AAB within the broader learning environment. Instructional guidance, assessment practices, the physical or digital learning environment, institutional policies, and time pressure all shape how AI tools are used and how their outputs are interpreted. For example, time pressure may amplify reliance on fluent AI outputs. By contrast, assessment practices that emphasize justification and reasoning may encourage more critical evaluation.

Supportive factors function as protective moderators that can attenuate AAB effects. These factors include metacognitive prompts, verification training, encouragement to explain, Socratic questioning, and teacher scaffolding. They interrupt the recursive AAB cycle by reintroducing effortful processing and self-monitoring (Flavell, 1979; Mehta et al., 2024). For this reason, supportive factors are particularly important for instructional design, because they represent concrete points of intervention.

At the learning-outcome level, these mechanisms shift learning from active knowledge construction toward surface-level processing. Learners may successfully reproduce correct answers or plausible explanations without developing the underlying conceptual structures required for transfer, adaptation, and flexible application of knowledge (Chi, 2009; Bauer et al., 2025). The decoupling of performance and understanding is the central pathway through which AAB shapes educational outcomes. This pattern is precisely what Fan et al. (2024) document empirically, providing converging evidence for the offloading and metacognitive components of AAB.

Domain Application: Research Methods Education

Research methods education provides a particularly informative context for examining AAB. Learning research methods requires multi-step analytical thinking, evaluation of evidence, and methodological justification (Mertens, 2019). These are demanding cognitive tasks. They depend on sustained engagement and careful self-monitoring. For this reason, the domain is especially sensitive to any reduction in cognitive effort or metacognitive control (Flavell, 1979; Chi, 2009). This makes research methods learning a useful setting for observing AAB in action.

Generative AI is increasingly used by students in research methods courses. Students turn to AI tools to obtain explanations of statistical concepts, to clarify research design choices, and to walk through data analysis procedures. These systems can provide structured and accessible explanations. However, the fluency and coherence of AI output may encourage learners to accept methodological interpretations without sufficient verification. This is particularly problematic in this domain. Methodological understanding does not depend only on reproducing procedures. It also depends on interpreting underlying assumptions and evaluating the validity of inferences (Shadish, Cook, & Campbell, 2002). When students accept fluent AI explanations without engaging with these deeper aspects, the consequences for learning are substantial.

Three domain-specific predictions follow from AAB:

- D1. AAB effects will be larger for conceptually integrative methodological topics, such as construct validity, sampling logic, and threats to internal validity, than for procedurally tractable topics, such as running a t-test in statistical software.
- D2. Students who rely on AI for methodological explanations will show inflated confidence when interpreting statistical outputs. However, they will show reduced ability to justify methodological decisions when prompted to articulate the underlying assumptions.
- D3. Iterative prompting interventions of the kind proposed by Mehta et al. (2024) will be more effective than warning-based interventions for reducing AAB in methods learning.

Together, these predictions show how AAB can generate domain-specific, testable hypotheses rather than remaining at the level of a general claim.

Discussion

Implications for Theory and Practice

The AAB framework has implications for teaching, assessment, and the design of AI learning tools. Each implication can be traced back to one of the four mechanisms in the framework. This means that the recommendations are not a general list of best practices. They are targeted interventions, each one designed to weaken a specific cognitive shortcut that AAB activates. We organize the implications below by the mechanism each one is designed to counter.

Visible reasoning to counter the fluency-to-accuracy illusion

Fluency is the most powerful input cue in AAB. Smooth, well-organized AI outputs are interpreted as accurate, even when the reasoning behind them is incomplete. This means that interventions should make reasoning processes visible rather than hidden. On the design side, AI systems should signal uncertainty and surface intermediate reasoning steps instead of presenting only polished final outputs (Kim et al., 2024). On the teaching side, students can be asked to evaluate AI outputs against alternative explanations, identify gaps in argument structure, or compare AI explanations with primary sources (Bloom, 1956; Chi, 2009). Both kinds of interventions force the reader to look past surface fluency and engage with the underlying reasoning.

Explanation- and justification-based assessment to counter cognitive offloading

Cognitive offloading is most valuable when assessment rewards correct answers alone. In that case, the student gains by handing the work over to the AI. However, assessment that emphasizes justification, reasoning steps, and transfer better reflects conceptual understanding (Roediger & Karpicke, 2006). This means that the value of offloading is reduced. When the assessment asks students to critique AI-generated explanations, reconstruct arguments, or apply methods to novel cases, cognitive effort is redirected toward active processing. In other words, the assessment design itself shapes how much offloading is rewarded.

Metacognitive scaffolding to counter reduced metacognitive monitoring

Reduced self-monitoring is one of the hardest mechanisms to address, because students often do not notice that their monitoring has been weakened. For this reason, instructional design needs to scaffold metacognitive engagement explicitly. Think-aloud protocols, guided self-questioning, and structured reflection on AI interactions can all support this kind of engagement (Flavell, 1979). Mehta et al. (2024) offer a concrete example with their iterative prompting framework. This framework requires students to interrogate AI outputs through follow-up questions. By doing so, it surfaces shortcomings in both the AI explanation and the student's own understanding at the same time.

Trust calibration training to counter the perceived authority heuristic

Authority heuristics produce automatic deference to confident, well-structured outputs. This deference can be reduced if learners are trained to calibrate their trust deliberately. Research on human-automation interaction shows that appropriate trust calibration depends on explicit verification habits, source-checking, and exposure to system errors in low-stakes contexts (Sheridan & Parasuraman, 2005; Parasuraman & Manzey, 2010). Educators can build these habits into instruction by giving students practice with AI outputs that contain visible errors, asking them to verify claims against trusted sources, and reflecting on cases where the AI was confidently wrong.

These four recommendations are not exhaustive. They illustrate how each AAB mechanism maps onto a distinct intervention target. Importantly, the four interventions are most effective when used together. Approaches that address only one mechanism, such as warning messages that target only authority cues, are likely to be less effective than coordinated interventions across all four mechanisms. This is consistent with the recursive nature of AAB described in Literature review section, where each mechanism reinforces the others. Breaking the cycle at one point may not be enough if the other mechanisms continue to operate.

Limitations and Future Research Directions

This study is subject to several limitations that future research should address. The most important limitation is that AAB, as a conceptual framework, has not yet been tested as an integrated whole. Several of its component mechanisms have already received empirical support, but the integrated pathway itself has not been examined directly.

A brief look at the existing evidence helps clarify the situation. Fan et al. (2024) provide evidence consistent with the cognitive offloading and metacognitive components of AAB. Kosmyna et al. (2025) provide neural evidence consistent with reduced engagement during AI use. Stadler et al. (2024) document the cognitive-ease-versus-quality trade-off. Finally, Mehta et al. (2024) and related work on IOED in AI contexts support the synthetic mastery outcome construct. Together, these studies show that the individual mechanisms in AAB are well documented. The empirical task is therefore not to test isolated mechanisms. Instead, the task is to test the integrated pathway that AAB specifies, in which these mechanisms operate together to produce the move from synthetic mastery toward false understanding.

Three research priorities follow from this. First, measurement work is needed. Researchers should develop an AAB-specific instrument that captures the configuration of fluency-based trust, cognitive offloading, and metacognitive miscalibration that AAB identifies. One promising starting point is the ETGAI-HE scale (Pandey et al., 2025), which already measures epistemic trust in generative AI in higher education. An AAB-specific instrument could build on or extend this scale.

Second, experimental designs are needed to isolate which AI input features matter most. Researchers should manipulate fluency, immediacy, and structured presentation independently, and then test whether AAB mediates the path from these input features to learning outcomes (Shadish, Cook, & Campbell, 2002). This kind of design allows the framework to be tested as a causal model rather than just a description.

Third, longitudinal studies are needed to examine how AAB effects develop over time. A key question is whether synthetic mastery, the in-the-moment subjective sense of understanding, stabilizes into durable false understanding over the course of a semester or a term. This is especially important in domains such as research methods, where the goal of instruction is transferred to new situations rather than reproduction of known answers.

Beyond these three priorities, future work should also test the boundary conditions specified in literature review section. These include the moderating roles of prior knowledge, AI literacy, task type, error salience, and the supportive factors that can attenuate AAB. Attention should be paid to the interaction between AAB and algorithm aversion (Dietvorst et al., 2015). The two constructs predict opposing patterns of trust under different conditions. AAB predicts inflated trust when AI outputs are fluent and confident. Algorithm aversion predicts reduced trust when errors are visible. Understanding when one dominates over the other is an important next step for theory and for practice.

Conclusion

AAB represents a meaningful shift in the cognitive dynamics of contemporary education. The shift is particularly visible in how learners evaluate and internalize AI-generated information. As generative AI systems become more integrated into learning environments, they shape epistemic judgment in distinctive ways. Fluency, immediacy, and perceived authority push learners toward reduced cognitive effort and greater reliance on external outputs. Recent empirical work confirms that these dynamics are not theoretical concerns. They are operating in real learning settings. Fan et al. (2024) document metacognitive laziness in students using ChatGPT. Kosmyna et al. (2025) measure neural disengagement during AI-supported essay writing. Stadler et al. (2024) describe the cognitive-ease and quality trade-off in AI-assisted research. Mehta et al. (2024) and related work track the illusion of explanatory depth in AI contexts. Together, these findings show that the cognitive consequences of AI use in education are observable and consistent across studies.

By introducing AAB, this paper provides a conceptual integration of these findings. The framework offers a unified pathway. It moves from AI input features, through cognitive mediators, to the learning outcomes of synthetic mastery and false understanding. The contribution of AAB is not the discovery of new biases. Instead, it lies in the articulation of how known mechanisms combine in conversational AI environments to produce a systematic miscalibration of epistemic trust. This means that the framework is integrative rather than novel at the level of any single mechanism. It explains why these mechanisms, which have been studied separately for decades, become especially powerful when they operate together in the context of a fluent and confident conversational AI.

It is also important to be clear about what AAB does not claim. AAB is not the claim that AI systems by themselves determine learning outcomes. Learning outcomes are shaped by how learners interpret and cognitively process AI-generated information. AI tools provide input. Cognition does the rest. The framework, therefore, places the locus of explanation inside the learner, not inside the technology.

These observations underscore a broader need. Educational approaches must promote critical evaluation, reasoning transparency, and metacognitive awareness in AI-supported learning environments. The goal of integrating AI into education should not be efficiency alone. Instead, it should be the preservation and strengthening of deep conceptual understanding and critical thinking. AAB offers a conceptual map for pursuing this goal. By identifying the specific mechanisms that lead to false understanding, it points to specific places where instruction, assessment, and AI system design can intervene. Future research can be built on this map by testing the integrated pathway, measuring AAB directly, and identifying the conditions under which protective factors are most effective.

References

Anderson-Cook, C. M. (2005). Review of *Experimental and quasi-experimental designs for generalized causal inference* by W. R. Shadish, T. D. Cook, and D. T. Campbell. *Journal of the American Statistical Association*. https://doi.org/10.1198/jasa.2005.s22

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Balta, N. (2026). False understanding in AI-assisted physics problem solving: A theoretical framework. *European Journal of Physics*. https://doi.org/10.1088/1361-6404/ae68a7

Bauer, E., Greiff, S., Graesser, A. C., Scheiter, K., & Sailer, M. (2025). Looking beyond the hype: Understanding the effects of AI on learning. *Educational Psychology Review, 37*(2), 45. https://doi.org/10.1007/s10648-025-10020-8

Belghith, Y., Mahdavi Goloujeh, A., Magerko, B., Long, D., Mcklin, T., & Roberts, J. (2024, May). Testing, socializing, exploring: Characterizing middle schoolers' approaches to and conceptions of ChatGPT. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-17). https://doi.org/10.1145/3613904.3642332

Bloom, B. (1956). *Taxonomy of educational objectives, Book 1*. Longmans, Green.

Cash, T. N., Oppenheimer, D. M., Christie, S., & Devgan, M. (2026). Quantifying uncert-AI-nty: Testing the accuracy of LLMs' confidence judgments. *Memory & Cognition, 54*(2), 375–400. https://doi.org/10.3758/s13421-025-01755-4

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*(1), 73–105. https://doi.org/10.1111/j.1756-8765.2008.01005.x

Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021, April). I think I get your point, AI! The illusion of explanatory depth in explainable AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (pp. 307-317). https://doi.org/10.1145/3397481.3450644

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*, 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., et al. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Efimova, E., & Nygren, T. (2026). Classroom discussions of social issues in the age of generative AI: Epistemic vigilance against bias and bullshit. *The Journal of Social Studies Research, 50*(2), 85–97. https://doi.org/10.1177/0885985x251382072

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2024). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology, 56*(2), 489–530. https://doi.org/10.1111/bjet.13544

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066x.34.10.906

Freire, P. (1970). *Pedagogy of the oppressed* (M. B. Ramos, Trans.). Continuum.

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and critical thinking. *Societies, 15*(1), 6. https://doi.org/10.3390/soc15010006

He, G., Kuiper, L., & Gadiraju, U. (2023). Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–18). https://doi.org/10.1145/3544548.3581025

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(5), 1191–1206. https://doi.org/10.1037/a0013025

Idowu, J. A., Koshiyama, A. S., & Treleaven, P. (2024). Investigating algorithmic bias in student progress monitoring. *Computers and Education: Artificial Intelligence, 7*, 100267. https://doi.org/10.1016/j.caeai.2024.100267

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences, 120*(11), e2208839120. https://doi.org/10.1073/pnas.2208839120

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

Jose, B., et al. (2025). Epistemic authority and generative AI in learning spaces: Rethinking knowledge in the algorithmic age. *Frontiers in Education, 10*, 1647687. https://doi.org/10.3389/feduc.2025.1647687

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kim, S. S., Liao, Q. V., Vorvoreanu, M., Ballard, S., & Vaughan, J. W. (2024). "I'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 822–835). https://doi.org/10.1145/3630106.3658941

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why does minimal guidance during instruction not work. *Educational Psychologist, 41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In *The Ethics of Artificial Intelligence in Education* (pp. 174–202). Routledge. https://doi.org/10.4324/9780429329067-10

Kosmyna, N., et al. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv*. https://arxiv.org/abs/2506.08872

Kumar, S., Mikayelyan, A., & Vorfolomeyeva, O. (2026). Fluency Illusion: A Review on Influence of ChatGPT in Classroom Settings. *Information*, *17*(3), 299. https://doi.org/10.3390/info17030299

Lee, H. P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025, April). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In Proceedings of the 2025 CHI conference on human factors in computing systems (pp. 1-22). https://doi.org/10.1145/3706598.3713778

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.

Lodge J. M. and Loble L (2026). Artificial intelligence, cognitive offloading and implications for education, University of Technology Sydney. https://doi.org/10.71741/4pyxmbnjaq.31302475

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607

Mehta, N., et al. (2024). Embracing the illusion of explanatory depth: A strategic framework for using iterative prompting for integrating large language models in healthcare education. *Medical Teacher, 47*(2). https://doi.org/10.1080/0142159X.2024.2382863

Mertens, D. M. (2019). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods*. Sage publications.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology, 67*(4), 371–378. https://doi.org/10.1037/h0040525

Mogavi, R. H., Deng, C., Kim, J. J., Zhou, P., Kwon, Y. D., Metwally, A. H. S., et al. (2024). ChatGPT in education: A blessing or a curse? A qualitative study. *Computers in Human Behavior: Artificial Humans, 2*(1), 100027. https://doi.org/10.1016/j.chbah.2023.100027

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 201–220). Erlbaum.

National Research Council. (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). National Academies Press. https://doi.org/10.17226/9853

Nguyen, T. N. T., Van Lai, N., & Nguyen, Q. T. (2024). Artificial intelligence in education: A case study on ChatGPT's influence on student learning behaviors. *Educational Process: International Journal, 13*(2), 105–121. https://doi.org/10.22521/edupij.2024.132.7

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12*(6), 237–241. https://doi.org/10.1016/j.tics.2008.02.014

O'Sullivan, J., Lowry, C., Woods, R., & Conlon, T. (2025). *Generative AI in higher education teaching and learning: National policy framework*. Higher Education Authority (Ireland). https://doi.org/10.82110/px37-mp48

Pandey, C. S., Mishra, P., Pandey, S. R., & Pandey, S. (2025). Epistemic trust in generative AI for higher education scale (ETGAI-HE scale). *AI & Society*. https://doi.org/10.1007/s00146-025-02566-6

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional self-regulation account. *Human Factors, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253. https://doi.org/10.1518/001872097779543886

Pitts, G., & Motamedi, S. (2025). Understanding human–AI trust in education. *Telematics and Informatics Reports*, 100270. https://doi.org/10.1016/j.teler.2025.100270

Reber, R., & Unkelbach, C. (2010). The epistemic status of processing fluency as source for judgments of truth. *Review of Philosophy and Psychology, 1*(4), 563–581. https://doi.org/10.1007/s13164-010-0039-7

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Sheridan, T. B., & Parasuraman, R. (2005). Human-automation interaction. *Reviews of Human Factors and Ergonomics, 1*(1), 89–129. https://doi.org/10.1518/155723405783703

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language, 25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: ChatGPT, learning, and argumentation in higher education. *Computers in Human Behavior*.

Tanchuk, N., et al. (2025). Personalized learning with AI tutors: Assessing and advancing epistemic trustworthiness. *Educational Theory*. https://doi.org/10.1111/edth.70009

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Corresponding Author Contact Information:

Author name: Koksal Mus

Department: Electrical and Computer Engineering Department

University, Country: Worcester Polytechnic Institute, Worcester, MA

Email: kmus@wpi.edu

ORCID: 0000-0001-9151-1989

Please Cite: Sony, S.D. & Mus, K. (2026). Algorithmic Authority Bias in AI-Assisted Learning: An Integrative Cognitive Framework. *International Educational Review, 4*(1), 75-93. https://doi.org/10.58693/ier.415

Conflict of Interest: The authors declare that they have no conflict of interest.

Data Availability Statement: Not applicable.

Ethics Statement: Ethical approval and informed consent were not required for this study because no human participants, animal subjects, or identifiable personal data were involved.

Author Contributions: **SDS:** Conceptualization, methodology, writing – review & editing, validation. **KM:** Conceptualization, methodology, writing – review & editing, validation. All authors approved the final version of the article.

Received: November 15, 2025 ▪ Accepted: March 12, 2026