

Article

From Plausibility to Auditability: Generative Systems, Synthetic Scholarship, and the Epistemic Risks of AI-Mediated Research

Reem Hashem ^{1,*}, Patricia Fidalgo ² , Yahia Mohammad Alramamneh ³, Eman S. Elkaleh ¹ , Farah El Zein ³ 
and Abdelaziz M. Hussien ³ 

¹ Data Analytics, Policy, and Leadership Division, Emirates College for Advanced Education, Abu Dhabi P.O. Box 126662, United Arab Emirates; eman.elkaleh@ecae.ac.ae

² Pedagogy, Technology and Teacher Education Division, Emirates College for Advanced Education, Abu Dhabi P.O. Box 126662, United Arab Emirates; pfidalgo@ecae.ac.ae

³ Language, Cognition and Growth Division, Emirates College for Advanced Education, Abu Dhabi P.O. Box 126662, United Arab Emirates; yahia.alramamneh@ecae.ac.ae (Y.M.A.); farah.elzein@ecae.ac.ae (F.E.Z.)

* Correspondence: reem.hashem@ecae.ac.ae

Abstract

GenAI is now capable of producing journal articles that look like human work, raising the question of whether the scholarly credibility might be misinterpreted as evidentiary validity. This research investigates AI-generated scholarship by shifting attention from the detection of AI-generated content to its critical appraisal and evaluation. To achieve this, a split-paper audit procedure was applied to a fully AI-generated mixed-methods manuscript that was independently reviewed by three university faculty members. Findings indicate that the model can simulate academic style and coherence with a high degree of surface plausibility; however, this superficial rigor was combined with ungrounded content, unsupported contextualization, selective reporting, and inconsistent methodological details. Reviewers' comments also highlighted the formulaic language and narratives that confirmed their prior expectations, and questioned whether prompt structures may be channeling outputs towards pre-existing assumptions and masking areas where evidence is scant. The study defines synthetic scholarship as an epistemic condition where plausibility may supplant verification, and suggests a transparent audit process to facilitate the assessment of AI-mediated scholarship. It argues that ensuring research integrity in the era of generative AI will rely on more than just disclosure of generative AI use; it will also require auditable trails to assist in examining, verifying, and substantiating claims.

Keywords: AI-mediated research; auditability; epistemic forensics; generative AI; methodological transparency; peer review; prompt framing; research integrity; scholarly publishing; synthetic scholarship



Academic Editor: Myint Swe Khine

Received: 9 May 2026

Revised: 23 June 2026

Accepted: 6 July 2026

Published: 14 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The world of academia is fast changing as large language models (LLMs) and other GenAI systems are being increasingly incorporated into research workflows. These tools are now employed to help with literature searches, drafting, translation, editing, data interpretation, peer-review preparation, and manuscript revision (Lund et al., 2023; Stokel-Walker, 2023). Their application has amplified a number of long-standing debates concerning authorship, responsibility, transparency, and the evidence supporting claims within academia

(Hosseini et al., 2023; Sahebi & Formosa, 2025). Publishing guidelines now increasingly acknowledge the possibility that AI-powered tools might have been used during the drafting phase of a manuscript; nevertheless, the humans who authored it bear the responsibility for its truth, integrity, and any disclosures relating to its production. For example, the International Committee of Medical Journal Editors (ICMJE) recommends that authors disclose the use of AI-based tools and describe the purpose and extent of their use in the manuscript (International Committee of Medical Journal Editors, n.d.). Elsevier also requires authors to disclose the use of AI tools in the writing of manuscripts and to specify the degree of human involvement (Elsevier, n.d.).

These developments suggest that the key issue is not just whether AI has been applied to the writing process; rather, the more critical question is whether scholarship that uses AI can be scrutinized, verified, and trusted. A manuscript may disclose its reliance on AI tools while still including citations, datasets, claims, or qualitative material that cannot be checked. Conversely, a fully human-authored manuscript may still require clearer evidence trails. This leads to the problem Sahebi and Formosa (2025) define as an overall epistemic trust issue when communication mediated by AI takes place—a problem where the reader needs to know when and to what extent it is appropriate to place trust in the other parties without falling into gullibility. The stakes in scholarly publishing are particularly high, as surface indicators of academic credibility are easy to be created without the corresponding grounding evidence.

These issues have resulted in a greater emphasis on the ways in which generative systems replicate academic practice. Outputs of large language models can mimic the style, structure, and rhetorical conventions of academic writing but with insufficient conceptual depth, evidentiary backing, or methodological transparency (Gao et al., 2023; Rudolph et al., 2023). The risk is particularly pronounced when a generative system is invited not just to help write but to undertake the full research process—designing a study; developing instruments; generating, analyzing, and interpreting data; and drafting a manuscript. Recent studies in generative-AI-enabled education similarly caution that such technologies may influence users' epistemic agency, critical thinking, and creativity as they take the output generated by these tools for granted (Pietarinen & Shi, 2026).

While the performance of LLMs is impressive when structured tests like professional exams are concerned, their logical reasoning abilities are often based on patterns and may be dependent on the way the question is phrased (Bommarito & Katz, 2022; Gilson et al., 2023; Hosseini et al., 2023; Kung et al., 2023). When it comes to academic writing tasks, generative AI models excel at summarizing and structuring narratives but are relatively weak at explaining, defining scope, substantiating theory, and validating sources (De Silva et al., 2024; Zimmermann et al., 2024). Generative models can therefore support technical aspects of scholarship, but human oversight remains necessary to safeguard accuracy, attribution, interpretation, and validity (Hosseini et al., 2023; White et al., 2023).

While previous work has examined AI as a writing assistant, learning support, or productivity tool, the present study shifts attention to AI as a simulated scholarly actor. It examines how a generative system constructs the appearance of a complete mixed-methods research study, including research instruments, simulated quantitative and qualitative data, analysis, interpretation, and manuscript writing. This paper moves beyond the issue of AI generated text fluency and investigates whether its assertions can be verified and trusted and whether its methods can be audited.

The study, therefore, addresses a pressing question for research integrity: how can scholars, reviewers, editors, and publishers assess research when the manifestations of academic credibility are algorithmically generated? To investigate this issue, the paper develops and implements a split-paper audit approach where an AI-produced manuscript is paired with

expert peer-review evaluations. The methodology permits the identification of markers of simulated rigor, such as bad citation practice, reporting bias, methodological inconsistency, invented data, synthetic qualitative voices, and prompt-induced confirmation bias.

The general objective of this work is to assess the performance of a generative AI system when prompted to write an entire mixed-methods research article, and to investigate whether simulated rigor can be detected by expert peer review. A further objective is to translate these markers into initial criteria for auditability that may be of assistance for editors, reviewers, and institutions in assessing work that is influenced by AI. The study is exploratory and illustrative rather than comparative or generalizable; it focuses on one AI-generated manuscript and three independent expert reviews to examine how scholarly form may be reproduced without sufficient evidentiary grounding. The study was guided by the following research questions:

- RQ1: How well can a generative system simulate the output of an entire mixed-methods research project, from instrument design and synthetic data generation to analysis, interpretation, and manuscript preparation?
- RQ2: What markers of simulated rigor can expert reviewers identify when evaluating an AI-generated research manuscript?
- RQ3: How does prompt framing shape the direction, coherence, and confirmatory tone of AI-generated scholarly claims?
- RQ4: What auditability standards are needed to help editors, reviewers, and institutions distinguish plausible scholarly form from verifiable scholarly evidence?

2. Literature Review

This literature review is organized around four areas relevant to the present study: the integration of generative AI into academic workflows; the simulation and homogenization of academic voice; audit approaches to AI-generated outputs; and emerging concerns around research integrity, disclosure, and verification in scholarly publishing. The review is narrative and conceptually focused rather than systematic. Its purpose is to identify the main strands of research that explain why auditability has become an urgent issue in AI-mediated scholarship and to show how the present study extends earlier work by examining a complete AI-generated mixed-methods manuscript through expert peer review.

2.1. AI in Academic Workflows

Scholars have responded with excitement and skepticism to the emergence of LLMs, in particular, the so-called advanced generative models. In terms of access and efficiency, these systems have facilitated the support of routine scholarly activities, including the generation of reading lists, the iteration of research questions, the development of outlines, and the preliminary drafting of manuscripts. Such manifestations of research may encourage wider participation in research, but issues of authorship, trustworthiness, and research integrity still dominate and moderate this enthusiasm (Lund et al., 2023).

In a widely discussed commentary in *Nature*, Stokel-Walker (2023) recounts situations in which a generative system was listed as a co-author on multiple published papers. The practice was universally condemned in the scientific community, where the ethical and intellectual arguments held that a machine cannot assume responsibility or accountability for errors associated with authorship. What was a matter of novelty at first sharply raised a deeper problem: the lack of, and need for, clear policies outlining a consensus on how to credit AI contributions to science.

Trust is a further complication. Haman and Školník (2023) conducted a systematic early inquiry, in which a generative system was instructed to write a literature review. While plausible existing citations were generated, many were fabricated, and several were

shown to be fictitious or contained typographical errors. These trends raise questions around the reliability of AI-generated academic writing and may risk undermining trust in the scientific record. The scholarly community is therefore moving into an area of promise and peril.

Recent publishing guidance has moved the debate from the question of “should AI tools be used” to “how such use should be disclosed and managed.” The Committee on Publication Ethics (COPE) considers AI an emerging domain of publication ethics that needs accountability and guidance for editors, reviewers, and authors ([Committee on Publication Ethics, n.d.](#)). The International Committee of Medical Journal Editors (ICMJE) requests that any AI-assisted technology used in the submitted work be disclosed, including details of how it was used in the cover letter and, as relevant, in the manuscript submitted ([International Committee of Medical Journal Editors, n.d.](#)). Elsevier declares that the policies on generative AI are designed to increase transparency and to support authors, reviewers, editors, readers, and contributors ([Elsevier, n.d.](#)). According to Springer Nature, AI cannot be given authorship and peer reviewers are requested not to submit manuscripts to generative AI platforms where confidentiality is not guaranteed ([Springer Nature, n.d.](#)). These policies indicate that authorship, disclosure, and human responsibility remain central. Still, they also reveal a further challenge: disclosure alone does not guarantee that the evidence, citations, data, or analytic procedures in a manuscript are trustworthy.

2.2. *The Simulated Scholar and the Homogenization of Academic Voice*

The idea of simulated agents with generative mechanisms has recently been on the rise. According to [Bozkurt \(2023\)](#), LLMs are autopoietic systems that recursively produce meaning through patterned linguistic production. Therefore, generative systems not only aid in academic writing but also reproduce the structural conventions of academic discourse through the production of texts that are similar to the organization and tone of academic texts.

However, as these systems are a prediction of the probability of linguistic patterns, the output tends to be more conventional and standardized. [Perkins \(2023\)](#) is concerned that this will further lead to the homogenization of academic voice, stifling creativity, critical discourse, and intellectual diversity. In an invited commentary, [Rudolph et al. \(2023\)](#) state that generative outputs are “a projection of accuracy” but “with no substantive content”. Empirical comparison studies offer further evidence to support this concern. [Amirjalili et al. \(2024\)](#) concluded that essays produced by AI were overall well-organized and persuasive, but they tended to be shallow, less creative, and cited poorer sources than those produced by humans in classroom settings.

2.3. *Auditing AI: Methodological Precedents and Reasoning Transparency*

There has been an increasing trend among researchers towards audit-style approaches to assess the capability and scope of generative systems. These studies compare the outputs of systems to human standards and offer an empirical approach to understanding reliability, reasoning transparency, and epistemic risk. In the field of law, [Bommarito and Katz \(2022\)](#) assessed a generative AI model on the Multistate Bar Examination. Even with the best prompts, the accuracy of the model was only slightly above 50%.

In medicine, [Kung et al. \(2023\)](#) and [Gilson et al. \(2023\)](#) assessed system performance on the United States Medical Licensing Examination, achieving accuracy rates of 43% to 64%. Reviewers observed that explanations were frequently clear and concise when the system was forced to justify its responses, even in cases of varying correctness. In a similar vein, [De Silva et al. \(2024\)](#) investigated generative classification of research articles and

found fairly strong categorical explanations but somewhat weaker accuracy in identifying scope and boundaries.

Others have warned that outputs that appear well-formed can be very misleading and need to be viewed with caution (Wu et al., 2023). Altogether, audit studies are somewhat mixed. Generative models can simulate some aspects of disciplinary competence and occasionally come up with compelling reasons, but the quality of their output is inconsistent, and they must be supervised by humans. Studies also show that generative reasoning is more pattern-based than conceptually based (Hosseini et al., 2023), and that language fluency can mask reasoning deficits (Rudolph et al., 2023).

2.4. From AI Detection to Epistemic Forensics

Early responses to AI-generated academic writing often focused on detection—determining whether a text was written by a human or a machine (Wu et al., 2023). However, detection is insufficient as research integrity strategy. A text may be AI-generated but carefully verified by human authors, or human-written but poorly evidenced and methodologically weak. The more important question is whether the knowledge claims in a manuscript are auditable. Epistemic forensics refers here to the systematic examination of the evidence trail behind a scholarly text, including the authenticity of data, the accuracy of citations, the transparency of prompts, the consistency of methods, and the relationship between claims and supporting evidence (Gao et al., 2023; Haman & Školník, 2023; Hosseini et al., 2023; White et al., 2023).

This shift is significant since generative systems can now generate not only fluent text but also credible tables, figures, statistical outputs, interview quotations, coding schemes, and methodological explanations. Such products may seem superficially cohesive but are actually disconnected with genuine empirical methods (Gao et al., 2023; Májovský et al., 2023). In this sense, editorial and peer review have to extend the evaluation of surface quality to include a consideration of whether the underlying evidence is accessible for scrutiny, capable of being reproduced, and/or subject to verification (Committee on Publication Ethics, n.d.; International Committee of Medical Journal Editors, n.d.; Springer Nature, n.d.).

2.5. Accuracy, Reliability, and Ethical Implications

Generative systems make academic writing and analysis more efficient, but their accuracy and reliability are still questioned. It has been shown that fake or non-existing citations could be found within generative outputs with a high level of linguistic plausibility (Haman & Školník, 2023). Also, Gao et al. (2023) reported that AI-generated abstracts were believable content-wise but lacked real empirical evidence. These characteristics may promote superficial acceptance of well-written outputs, even among expert reviewers, without sufficient verification. When claims are not the product of open evidentiary processes, uncertainty and the effect of partial corroboration might undermine scholarly validity.

These issues raise significant ethical concerns. Fake or unverifiable citations can result in misleading or unsupported scholarship, and this can call into question the reliability of the scientific literature (Hosseini et al., 2023). There is emerging evidence for the use of AI generated text in academic manuscripts and grant proposals, raising concerns about transparency and accountability as well as potential propagation of pseudoscientific ideas when verification is not robust (Wu et al., 2023; Chen et al., 2023).

A proposed solution is structured prompting with critical oversight. White et al. (2023) propose that good prompts (e.g., which include questions for verifiable references and explicit reasoning) can reduce hallucination and increase transparency but are not able to completely remove such risks. In the literature, there are three emergent and consistent

principles. First, the ultimate accountability for accuracy should rely on the human author (Hosseini et al., 2023). Second, solid verification mechanisms and procedures are required to assess generative outcomes (Haman & Školník, 2023; Gao et al., 2023). Third, prompt design significantly impacts the quality of the outputs, although prompt engineering cannot replace human responsibility or critical assessment (White et al., 2023).

Generative systems are known for their high processing speed in producing text summarization and outlining or synthesizing large amounts of information. Comparative research suggests that these systems are frequently superior to humans when it comes to working with procedural text and document navigation. Efficiency, however, does not necessarily guarantee analytic rigor. A case study by Zimmermann et al. (2024) reported that, despite time and resource savings from generative assistance, the accuracy of system-generated reviews remained moderate. The analysis further noted a trend for such simplifications, redacting of good quality material, and a lack of adequate contextual understanding.

In parallel, similar concerns arose about the impact of generative efficiency on the depth of scholarship. Perkins (2023) claims that reliance on automated text generation might encourage surface reading, as outputs look comprehensive but lack critical engagement and interpretive complexity. Linguistically coherent but superficial summaries may induce a false sense of completeness, which could potentially be used to hide gaps in logic and evidence.

2.6. User Perspectives and Educational Impacts

New findings suggest greater use of generative systems in student learning activities. Alnaim (2024) states that generative tools are increasingly being used by higher education students in essay writing and assignment preparation. Students were able to identify some perceived benefits, including motivation, positive feedback, and assistance in organizing their work. The findings highlight that generative systems may foster access and serve as a learning scaffold when implemented in a supportive and mediational approach.

Despite the benefits of generative systems in academia, concerns remain about their potential displacement of core academic skills. Some students depend solely on generative systems to complete their assignments, reading, thinking and analyzing—the very skills higher education exists to teach. Concerns have been raised by Perkins (2023) and Amirjalili et al. (2024) that the mass adoption of automated text generation might result in a homogenization of the student voice and a diminution of higher-order thinking skills. Comparable observations for academic staff are also reported, with generation tools being used for editing, literature searches, and preliminary writing (Liu et al., 2025; Maistry & Singh, 2025). While many acknowledge the positive aspects of greater efficiency and timesaving, unease lingers over plagiarism, muddy authorship, and the upholding of scholarly rigor.

The studies reviewed above show important progress in understanding AI-generated academic writing. Still, they also reveal a recurring limitation, as much of the existing work examines isolated outputs, such as abstracts, essays, citations, or task-based responses, and not the full simulation of an empirical research manuscript.

2.7. Synthesis: Toward Auditability in AI-Mediated Scholarship

The literature shows that generative systems can support scholarly and educational work by improving efficiency, assisting with organization, helping users draft or revise academic texts, and supporting some aspects of learning and feedback (Alnaim, 2024; Lund et al., 2023; Maistry & Singh, 2025; Zimmermann et al., 2024). At the same time, these systems can generate fabricated citations, unsupported claims, formulaic arguments, and apparently coherent but unverifiable research narratives (Gao et al., 2023; Haman & Školník, 2023; Májovský et al., 2023; Rudolph et al., 2023). The key limitation in the existing

literature is that many studies identify risks at the level of discrete outputs. Still, fewer examine how these risks appear when a generative system is asked to simulate an entire empirical research process. The problem is therefore not simply whether generative AI is used, but whether AI-mediated scholarship remains accountable to the standards of evidence, transparency, and critical reasoning that define credible research (Hosseini et al., 2023; Perkins, 2023; White et al., 2023).

To clarify how the present study builds on, but also moves beyond, previous research, Table 1 summarizes key studies, their main findings, their limitations, and their relevance to the proposed auditability approach.

Table 1. Prior research, limitations, and relevance to the present study.

Study/Source	Focus	Key Finding	Limitation or Gap	Relevance to the Present Study
Haman and Školník (2023)	Use of ChatGPT to conduct a literature review	The system produced some credible sources but also cited fabricated or inaccurate citations.	Focused on literature review generation rather than a complete empirical manuscript.	Supports the need for citation integrity as an auditability criterion.
Gao et al. (2023)	AI-generated scientific abstracts	AI-generated abstracts appeared believable to reviewers but lacked real empirical grounding.	Focused on abstracts rather than full research articles.	Shows how plausibility can be mistaken for evidentiary validity.
Májovský et al. (2023)	AI-generated medical articles	AI generated fraudulent but authentic-looking scientific articles.	Demonstrated the risk but did not offer a detailed reviewer-facing audit framework.	Supports the need to move from detection to verification and auditability.
Amirjalili et al. (2024)	Comparison of AI-generated and human academic writing	AI-generated texts were organized and persuasive but often shallow and less original.	Focused on academic writing quality rather than research process simulation.	Informs the discussion of homogenized academic voice and reduced critical diversity.
White et al. (2023)	Prompt engineering and prompt patterns	Structured prompts improved outputs but could not remove the need for human judgment and verification.	Did not examine how prompt design shapes complete research manuscripts.	Supports the analysis of prompt framing and confirmatory alignment.
Wu et al. (2023)	Detection of LLM-generated text	Detection remained methodologically difficult and imperfect.	Detection did not address whether claims, data, and citations were auditable.	Supports the study's shift from detection to auditability.
Pietarinen and Shi (2026)	Generative AI-mediated education, manipulation, and epistemic agency	Generative AI influenced users' critical thinking, agency, and interpretation of information.	Focused on educational mediation rather than scholarly publishing.	Supports the need to protect critical judgment in AI-mediated knowledge work.
Sahebi and Formosa (2025)	AI-mediated communication and epistemic trust	AI-mediated communication created new problems of trust, credibility, and justified reliance.	Did not focus specifically on manuscript peer-review or empirical research simulation.	Strengthens the study's framing of auditability as an epistemic trust issue.
This study	Split-paper peer-review audit of a complete AI-generated mixed-methods manuscript	Showed how a generative system can reproduce scholarly form while producing weakly auditable evidence, synthetic qualitative voices, and prompt-shaped coherence.	Exploratory single-case audit based on one generated manuscript and three expert reviews.	Extends previous work by examining a full simulated empirical manuscript and proposing preliminary auditability criteria for editors, reviewers, and institutions.

Three aspects are particularly salient for the present study. First, generative systems can reproduce the form of scholarship without necessarily reproducing its evidentiary foundations (Amirjalili et al., 2024; Bender et al., 2021; Gao et al., 2023). Second, prompt design can influence the direction and tone of outputs, promoting coherence, convergence, and confirmation while deterring engagement with contradiction, doubt, or alternative

interpretations (White et al., 2023). Third, current policy guidance is increasingly focused on disclosure and the responsibility of the individual, but disclosure itself does not tell us whether the citations, data, methods, or claims on which it is based can be trusted (Committee on Publication Ethics, n.d.; Elsevier, n.d.; International Committee of Medical Journal Editors, n.d.; Springer Nature, n.d.). These considerations indicate that practical audit procedures should look beyond detection of AI-generated text to determining whether the evidence underpinning scholarly claims can be audited.

The current study contributes to this gap by subjecting one AI-produced mixed-methods manuscript to a split-paper peer-review audit. We do not seek to make generalizable statements about all generative AI systems or all AI-mediated scholarship. Instead, the analysis adopts a zoomed-in exemplary case approach to analyze a full research manuscript and expose how simulated rigor manifests itself in evidence, method, citation practice, qualitative voice, and prompt-shaped interpretation and how expert reviewers uncover these weaknesses. Based on this, a preliminary set of auditability requirements is suggested that can be further validated in forthcoming studies.

This body of literature informs the methodological design of the study; if generative systems can produce surface features of scholarship, then peer review needs to be assisted by tools that investigate the traceability and verifiability of the evidence undergirding those features.

3. Methodology

3.1. Design

We used a qualitative split-paper peer-review audit design to investigate how a generative AI system performs when it is a simulated scholar. This design was chosen because it allows for a dual evaluation of the generated manuscript: as academic writing and as an auditable scholarly artifact. In this setup, the draft produced by the AI is the object of analysis, and the comments of the peer reviewers (experts) serve as ‘evidence’ to identify simulated rigor, evidentiary weakness, methodological opacity, and prompt-shaped coherence. The research is exploratory and descriptive, not comparative; it considers a single AI-generated paper and three separate expert reviews.

The study was carried out in two successive phases:

1. The creation of the simulated research manuscript: The generative AI model was prompted to generate a full mixed-methods research manuscript, encompassing the research problem, framing of the literature, instruments, simulated qualitative and quantitative data, analysis, results, and discussion.
2. Critical expert review audit: The output manuscript was independently critiqued by three university faculty with expertise in mixed-methods research design, educational leadership, and AI ethics/research integrity. Reviewers assessed the manuscript according to traditional scholarly standards of novelty, methodological soundness, evidential reliability, ethical transparency, academic writing, and trustworthiness.

The resulting artifact was presented in split-paper form. The unedited AI-generated manuscript was on one side, and the reviewer comments (with the relevant manuscript sections) were on the other side. The layout enables direct juxtaposition of AI-generated scholarly claims and expert human assessment. The split-page format thus served as a working audit record, exposing where the manuscript looked believable in terms of the “scholarly form” but was undermined by weak evidence, method, or interpretation. None of the reviewers’ identities of the manuscript were disclosed. We only report broad information relating to the subject areas of expertise. The use of AI generation was transparently disclosed as a component of the study design; nevertheless, reviewers did not have access to the prompt archive, generation workflow, or researcher notes. Doing so ensured that

their comments on the task would be focused on the intellectual quality and auditability of the produced manuscript rather than on the prompting procedure itself. Figure 1 illustrates the split-paper peer-review audit procedure adopted in this study. The procedure began with the generation of a complete AI-produced mixed-methods manuscript and proceeded through independent expert review, section-by-section mapping, thematic synthesis, prompt-framing analysis, and the development of preliminary auditability criteria.

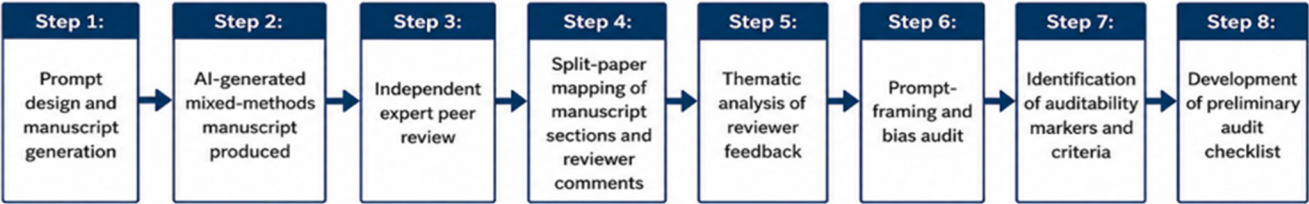


Figure 1. Split-paper peer-review audit procedure.

3.1.1. Stage 1: AI-Generated Manuscript

In Stage 1, the research team generated a full mixed-methods manuscript using ChatGPT 4.5. No other generative AI tool was employed to produce the manuscript analyzed in this study. The precise model version and system date were not stored in the audit trail, which is recognized as a limitation as version controls of models might have impact on output quality, hallucination patterns, and reproducibility. The choice of one system was in line with the exploratory single-case audit study design, which intended to focus on one produced scholarly artifact for an in-depth examination rather than comparing the performances of a number of models. The simulated study did not involve any human participants; all data, quotes from participants, statistical results, and qualitative themes were produced by the system. The produced manuscript was centered around a fictitious K–12 education research in the United Arab Emirates—Emotional Intelligence and School Leadership in the UAE: Driving Digital Transformation and Change Adoption in Education.

The model was instructed to:

- Formulate research questions and design a mixed-methods study;
- Build research instruments (survey and interview /focus-group protocols);
- Generate quantitative and qualitative datasets;
- Conduct statistical and thematic analyses;
- Generate a full-length research paper, including tables, figures, and conclusion.

To facilitate transparency, the prompts, instructions, outputs, and intermediate drafts utilized in manuscript construction were retained by the research team. The prompt archive was employed to analyze the effect of input phrasing on the selection of constructs, methodological assertions, simulated results, and interpretive tone by the system.

Prompting was considered as a means of procedure and as a unit of analysis. From a procedural perspective, prompts instruct the system to generate specific components of a research manuscript (e.g., study rationale, research questions, instruments, simulated datasets, statistical results, qualitative themes, discussion, and conclusion). From an analytic standpoint, the prompts enabled the researchers to investigate how linguistic signals like “rigor,” “integration,” “triangulation,” “limitations,” and “theoretically aligned findings” impacted the produced output.

The prompt sequence is summarized in Table 2 to clarify the prompting process. The table shows the major prompting stages, the aim of each stage, the content focus, and the output type. The table provides a transparent account of the prompting workflow used to generate the manuscript.

Table 2. Prompting procedure and generated outputs.

Prompt Phase	Purpose	Content Focus	Output Generated
Study framing	To establish the topic, setting, and research problem	Emotional intelligence, school leadership, digital transformation, UAE K–12 education	Research problem, study aim, research questions, and hypotheses
Methodological design	To generate the overall research design	Mixed-methods design, quantitative and qualitative strands, sampling, and instruments	Methodology section and study procedures
Instrument development	To create simulated research tools	Survey constructs, interview / focus-group protocols, leadership, and teacher variables	Survey measures, interview questions, and focus-group prompts
Data generation	To produce simulated empirical material	Quantitative data, qualitative quotations, school-level context, and participant-style responses	Statistical results, participant quotations, and qualitative themes
Analysis and integration	To generate findings and connect data strands	Mediation, moderation, thematic analysis, and mixed-methods integration	Results tables, integrated findings, and explanatory claims
Manuscript drafting	To produce the full article	Abstract, introduction, literature review, methods, results, discussion, and conclusion	Complete AI-generated manuscript
Refinement prompts	To improve coherence and scholarly form	Rigor, integration, limitations, theoretical alignment, and implications	More polished and internally coherent manuscript sections

The following details describe the design of the AI-generated manuscript, not an empirical study conducted with real participants:

- Simulated quantitative phase: The generated manuscript described surveys of teachers and school leaders from approximately 50–60 schools, with reported sample sizes ranging from 800 to 1200 participants. These figures were generated by the AI system and were later examined as part of the audit because they were reported as ranges rather than exact sample sizes.
- Simulated qualitative phase: The generated manuscript described semi-structured interviews with 25–30 leaders and 8–10 focus groups, each consisting of 6–8 teachers. These interviews and focus groups were not conducted; the quotations and themes were synthetic outputs generated by the AI system.

The generated manuscript referenced the following measurement instruments:

1. Wong & Law Emotional Intelligence Scale (Wong & Law, 2002).
2. Organizational Trust Inventory—Short Form.
3. Edmondson's (1999) Psychological Safety Scale.
4. Organizational Readiness for Implementing Change (Weiner, 2009).
5. Unified Theory of Acceptance and Use of Technology (Venkatesh et al., 2003).
6. Technostress Scale (Tarafdar et al., 2007).
7. A newly generated Regulatory Pressure and Cultural Diversity Scale.

In the produced paper, the integration of both data streams using qualitative and quantitative methods was done through theme identification and descriptive reasoning about the data. The analysis processes used in the study, such as multilevel structural equation modeling, mediation, moderation analysis, and thematic analysis, were also described. The above claims have been assessed in Section 4 (Results) to evaluate the sufficiency of the methodology presented in the paper.

3.1.2. Stage 2: Peer Review

The peer-review audit was conducted by three independent university faculty reviewers. Reviewers were selected because their expertise collectively covered mixed-methods research design, educational leadership, and AI ethics/research integrity. To preserve anonymity, the manuscript reports only broad professional information rather than names, affiliations, or identifiable demographic details. Each reviewer had experience evaluating scholarly work in higher education. The reviewers worked independently and did not discuss the manuscript with one another before submitting their written reports. Each reviewer was provided with the full AI-generated manuscript but not the prompt archive, generation workflow, or researcher notes. Reviewers were instructed to assess the manuscript by normal scholarly standards rather than attempting to recreate the prompting process. The review process consisted of five major components, which included the following:

- Originality and scholarly contribution;
- Methodological soundness and transparency;
- Ethical considerations;
- Academic tone and authorial voice;
- Overall credibility.

Reviewer comments were compiled and mapped onto the corresponding sections of the AI-generated manuscript in the split-paper format. This enabled the research team to contrast particular AI-generated claims, methods, results, and interpretations with responses from expert reviewers. The mapping also revealed common weaknesses reported by reviewers such as citation holes, methodological opaqueness, invented evidence, reporting by selection, and formulaic academic voice.

The assessment of simulated scholarly output of the generative system was done via structured expert peer review. The reviewers had to judge the paper according to the usual standards of scholarship. There were no numerical scores or automated evaluation instruments involved in the process. The reviewers were not instructed to identify if the submission was AI-generated, either. This decision reflected the objective of the research, to explore if the standard peer-review decision-making could detect flaws in the underpinning of evidence, methodological transparency, and scholarly credibility when the manuscript appeared structurally coherent and academically fluent.

3.2. Data Analysis Strategy

The analysis of the three peer-review reports followed a three-step process.

1. Section-by-section critique: The manuscript produced by the AI was reviewed one section at a time. Reviewer comments were matched to relevant manuscript sections to indicate the specific claims, methods, findings, or interpretations the reviewers considered to be weak, unsupported, unclear, or not sufficiently auditable.
2. Thematic and synthetic analysis: Two researchers independently analyzed the three peer-review reports and developed initial codes concerning citation support, methodological transparency, evidentiary credibility, prompt framing, selective reporting, and academic voice. The researchers then discussed their coding, clustered similar codes to potential themes, and refined these themes through discussion. Discrepancies were settled by consensus rather than statistical inter-rater reliability. This approach aligned with the interpretive intent of reflexive thematic analysis, where the goal was not to generate a numeric reliability estimate but rather to construct a meaningful and defensible thematic interpretation (Braun & Clarke, 2006).
3. Framing of prompts and bias audit: The research team analyzed the prompt archive to assess how the wording of the inputs affected the inclusion of constructs in the paper,

the simulated results' direction, the existence or nonexistence of opposing evidence, and the quant/qual claim concordance. This phase considered prompts as active epistemological inputs rather than passive instruction sets.

3.3. Auditability Criteria

The analysis was governed by a number of auditability criteria that were derived from three sources: previous work on AI-generated weaknesses in scholarly work, ethical principles around publication and disclosure of AI, and through an inductive analysis of all three reports in the peer-review process. The analysis was governed by a number of auditability criteria that were derived from three sources: previous work on AI-generated weaknesses in scholarly work, ethical principles around publication and disclosure of AI, and through an inductive analysis of all three reports in the peer-review process. As such, the criteria are knowledge-informed and exploratory rather than subject to assessment of psychometric properties or other forms of formal validation. They served as a guide in interpreting the shortcomings of the produced manuscript and in organizing the suggested checklist for editors and reviewers.

The criteria were operationalized as follows.

Citation integrity concerns whether cited sources actually exist, are accurately depicted, and directly support the associated claims. In this study, citation strength was evaluated based on reviewer comments that detected the absence of references or presence of weak, fabricated, or unsupported references.

Transparency on data generation is a statement on whether the data are real, simulated, synthetic or AI-generated, in the manuscript information. In this work, the produced paper was analyzed on the basis of whether it explicitly separated simulated from empirical evidence.

Sampling consistency is about reporting sample size, participant groups, and inclusion/exclusion criteria consistently and accurately. Uncertain participant classifications, samples sizes that varied, or numbers that were not consistent were considered as matters of auditability.

Instrument transparency is defined as whether instruments are named, cited, described, and accompanied by information on instrument validity or reliability. Without any information about the items, the response scales, or the original citations, the instruments were considered not auditable.

Traceability of qualitative evidence is the extent to which you can follow the evidence from participant quotations back to transcripts, field notes, anonymized excerpts, or some other form of secure audit trail. In the resulting manuscript, the quotations from participants' style were analyzed as synthetic voices, as no actual interviews or focus groups had taken place.

Alignment of the claim and evidence refers to whether claims are directly supported by the data, citations, or analysis in the article. Claims were auditable with difficulty if they were plausible sounding, but they did not have a visible evidentiary support.

Selective reporting is concerned with whether null, weak, contradictory, or unanticipated results are reported together with positive or confirmatory results. The lack of disconfirming evidence was considered as an indicator of over-confirmation in reporting.

Prompt transparency is about whether the AI system, the intended purpose, the prompting strategy, and the level of human interaction are made known. This standard is particularly significant as the wording of prompts can affect the constructs, results and interpretative trajectory of the outputs generated by AI.

These guidelines were not applied as a numerical scoring system. Instead, they served as interpretive guidelines for identifying where the output of the system produced a manuscript that appeared to be rigorous but that was unsupported by adequate evidence.

As the criteria have not been formally validated, they are intended as initial auditability criteria that should be further evaluated across other manuscripts, disciplines, reviewers, and generative systems.

3.4. Rigor and Reporting Standards

Rigor was underpinned by independent expert review, systematic comparison of reviewer comments, thematic analysis of recurrent issues, and retention of the prompt and output audit trail. The analysis of one model and one generated manuscript was not intended to statistically generalize. Instead, it adopted an exploratory audit methodology to determine transferable risks and evaluative indicators for AI-mediated scholarship.

The authors did not employ any independent master key to check the reviewers' comments because the study was focused on understanding how experts evaluated a created scholarly artifact, not on comparing their decisions with a predefined evaluative standard. The analysis instead depended on reviewer convergence as its main means of validation. Similar concerns that were independently identified by multiple reviewers were taken as stronger signals to potential weaknesses of auditability, since repetitive remarks denoted a higher degree of agreement amongst experts. Such a flexible reporting approach is in line with the nature of qualitative enquiry in audit, where the aim is to capture patterns of expert judgements as opposed to assessing these judgements for accuracy against an external benchmark. However, future research might enhance methodological rigor by formulating and testing formal audit protocols or evaluative rubrics that could be applied systematically to larger sets of AI-produced manuscripts.

The professional background of the reviewers in research methodology, educational leadership, and AI ethics/research integrity served as the foundation for the evaluation of coherence, evidentiary support, methodological transparency, and scholarly credibility. The limitations of this approach are discussed in the Limitations section, especially the consideration of a generated manuscript, a generative system, and a number of expert reviewers.

4. Results

The three peer-review reports were synthesized into a combined evaluation of the manuscript generated by the AI. Throughout this exploratory single-case audit, a pattern became clear: at the level of form, the manuscript was believable as an academic document, but as an evidence document, it was thin. The results are presented through five layers: section-by-section critique; thematic convergence and divergence between the themes; prompt-framing and directionality bias; markers of auditability identified by reviewers; and, ultimately, synthesis vis-a-vis the research questions.

4.1. Section-by-Section Critique

A section-by-section commentary on the manuscript appears in Table 3. It juxtaposes the reviewers' comments with a selection of extracts from the generated manuscript, enabling the reader to identify the exact location in the manuscript where the structure of a traditional mixed-methods paper was replicated and where weaknesses related to evidence, citation support, methodological transparency and reporting balance were detected by reviewers.

In Table 3, it could be observed that the AI-generated paper reached a high level of structural and rhetorical plausibility, which was almost the same as the academic writing style in general. So, its failings were not in its shape but in the evidence and methodological and interpretive practices that constituted the shape. Its architecture was said to be one of the most compelling features. The trouble was that this arrangement was giving a semblance of unity that was not anchored in verifiable facts. The analytic procedure was

not easily auditable due to methodological opacity. Consequently, many of the findings were not subjected to rigorous interrogation and were instead granted the credibility expected of scholarly output. The comments from the reviewers consequently highlight a difference between apparent and actual rigor that is at the very core of this paper; the text created the impression of a genuine academic paper without the evidentiary basis, methodological detail and caution of interpretation that makes scholarship reliable. At the section level, the main findings were as follows:

- The Abstract conformed to the conventional reporting format of a research summary in this discipline with the following categories: background, purpose, design, sample, and findings. However, it has given simulated results as if they are empirically proven, which offered the premature validity to the readers.
- The Introduction and Literature Review sections displayed a coherent thematic flow, despite relying on sweeping generalizations and unsupported contextual statements with minimal citation backing. Reviewers commented that a few statements regarding the UAE context, digital transformation, and emotional intelligence were only stated rather than substantiated.
- The Conceptual Foundations section presented a number of pertinent theories but did not always include definitions or provide rationale for crucial variables (i.e., regulatory pressure, cultural diversity, psychological safety, technostress buffering) that would be utilized subsequently in the study. This resulted in a disconnection between conceptual framing and analytic claims.
- The greatest limitations were noted in the Methodology and Results sections. Reviewers noted inconsistent reporting of participants sample sizes, poor description of measures, inadequate details of quantitative analysis, unclear qualitative procedures, and limited demonstration of real integration in mixed methods.
- The resulting manuscript asserted that multilevel structural equation modeling facilitated the examination of direct, mediated, and moderated relationships among leaders' emotional intelligence, teacher trust, psychological safety, and readiness for change. However, none of the reports provided model-fit indices, information on estimation methods, or details about the measurement-invariance testing. Sample sizes were also noted as approximate ranges (e.g., 800–1200 teachers), and the seven instruments were identified without precise item detail, evidence of reliability, or clarity of source.
- The qualitative phase purported to apply [Braun and Clarke's \(2006\)](#) thematic analysis but did not report a coding framework, theme hierarchy, or trustworthiness procedures. Quotes from the qualitative findings were vivid and relevant, but it was not possible to trace these back to transcripts, field notes, or actual participants. The Results section also made hypothesis-supporting relations conspicuous by hiding null, weak, and contradictory results.
- The Discussion and Conclusion sections were relatively short and not adequately linked to the literature. Reviewers also pointed out that implications for practice and policy were limited.

These findings indicate simulated rigor, where the thickness of the methodological language, multilevel modeling, mediation, moderation, thematic analysis, and integration was that of high-level research, but the procedural detail needed to render these claims auditable was not supplied. In this sense, the risk presented in this article did not just pertain to the errors of fact; the manuscript presented the façade of being methodologically sound.

Table 3. Split-design review.

Paper Section/Extract	Reviewer Comment(s)
Abstract	
<i>“The purpose of this study is to investigate how school leaders’ emotional intelligence (EI) influences digital transformation and change adoption in K-12 schools in the United Arab Emirates (UAE) ... An explanatory sequential mixed-methods design was employed ... Findings reveal that leaders’ EI significantly predicted teacher readiness ...”</i>	R1: Abstract includes concise statements about context, purpose, and significance. Provides overview of research design, phases, and sample. Ends with a summary of findings.
Introduction	
<i>“Evidence consistently shows that the primary barriers to reform are not technical limitations, but human and organisational challenges—teacher resistance, technostress, fragile trust, and limited readiness to embrace change.”</i>	R1: Solid assertions about barriers and limitations cannot be substantiated with citations. Paragraph two needs a clearer connection between barriers and EI. Studies claiming gaps in the literature in paragraph three must be supported with citations. H2 and H3 contributing factors—trust, safety, regulation, diversity—should be introduced earlier to help frame your study’s contribution and improve the structure of the paper. The final roadmap paragraph could be removed and replaced with a restatement of purpose and significance.
<i>“... Much of the educational leadership literature emphasises general outcomes such as school climate or teacher satisfaction, with relatively little attention to the specific dynamics of digital transformation ... research designs have been predominantly cross-sectional and reliant on self-report data.”</i>	R2: A number of claims are not backed up by references. Recent research should be acknowledged. Unsupported claims regarding the use of self-report and cross-sectional designs.
<i>“... three hypotheses are tested: (H1) leaders with higher EI ... (H2) trust and psychological safety ... (H3) regulatory pressure and cultural diversity ...”</i>	
Literature Review	
<i>“In the United Arab Emirates (UAE), a rapidly evolving K-12 sector places principals at the forefront of ambitious digital transformation initiatives and continuous policy reforms. Leaders operate within top-down modernisation mandates and highly multicultural school communities. Ensuring that principals possess not only technical and instructional competence but also strong emotional intelligence (EI) has therefore become increasingly salient.”</i>	R1: The section is too lengthy (7+ pages) and needs to be shorter. It started with what I think is information that is not supported by the literature. The review aim, search process, and analytic scheme should be clearly articulated to establish the context for the findings/discussion. Findings/discussion needs to be shortened as well so that it can serve as a basis for the methods section.
<i>“... The UAE is characterised by ambitious reform cycles, inspection-driven accountability, and extreme staff cultural diversity. Local empirical work shows EI’s salience.”</i>	R2: Many claims unsubstantiated (e.g., UAE context, ambitious reforms, cross-sectional critique). Missing references. Repetition of EI models; some implications are better placed in the discussion. Citation issues such as missing a page number). Positive: strong discussion of EI alongside other leadership theories.
<i>“... many studies are cross-sectional and single-source, constraining causal inference and inviting common-method bias.”</i>	R3: Sections begin like “announcements” (non-academic style). Strong statements lack citations.
Conceptual Foundations	
<i>“Early definitions by Salovey and Mayer (1990) described EI as the ability to monitor one’s own and others’ feelings ... Mayer, Salovey, and Caruso (2016) delineated EI into four branches ... In contrast to this ability-centered view, mixed or competency-based models emerged during the 1990s ...”</i>	R1: Key concepts need to be operationally defined. For example, what do the following variables really mean: teacher readiness, trust and psychological safety, regulatory pressure, cultural diversity, and technostress buffering?
<i>“... Across frameworks—transformational, instructional, distributed, and authentic leadership—EI functions as an enabling substrate rather than a competing style.”</i>	R3: Many citations are missing. Some parts read like a discussion and should be moved there. Regulatory pressure (a moderator in the study) is not covered in this section. Cultural diversity is mentioned randomly, without systematic treatment.
<i>“... Educational change is rarely purely technical; it is saturated with fear, resistance, hope, and optimism. Fullan (2001) described the culture of change as ‘rife with anxiety, stress, and ambiguity’ ...”</i>	
<i>“... In a system like the UAE—where digital transformation is tightly coupled to national development goals and schools are microcosms of global diversity—EI becomes indispensable ...”</i>	

Table 3. Cont.

Paper Section/Extract	Reviewer Comment(s)
Methodology & Results <i>“This study employed an explanatory sequential mixed-methods design (QUAN → QUAL) ... Phase 1 involved large-scale quantitative surveys ... Phase 2 involved qualitative interviews and focus groups ... Integration occurred at the interpretation stage through joint displays and meta-inferences ...”</i> <i>“The study was conducted in K-12 schools across the United Arab Emirates, including Abu Dhabi, Dubai, Sharjah, and two additional emirates.”</i> <i>“Measures included EI (WLEIS), trust, psychological safety, readiness, UTAUT adoption constructs, technostress, and contextual moderators ...”</i> <i>“Quantitative analysis indicated that leader EI was positively associated with teacher trust ($\beta = 0.31, p < 0.001$) and psychological safety ($\beta = 0.26, p < 0.001$) ... Interview data reinforced these findings ...”</i>	R1: The key terminology (teacher readiness, trust, safety, regulatory pressure, cultural diversity, technostress) need definitions, in addition to the design, phases, and hypotheses are concise and flowcharts clear. Context: why are only some emirates named? Include summary of AI curriculum. Sampling: quantitative phase should explicitly mention “survey.” Unclear what the specific reason for these measures was, inconsistent presented numbers, and it is unclear why citation of some scales is missing. Phase 2: not enough description of interview / focus group protocols, aims, and measures. Data analysis: more detail for SEM methods, coding, member checking, and integration of strands was needed. Suggest removing “timeline/deliverables.” Results: report negative/weak findings too; clarify correlation tables; provide qualitative themes with subthemes and distinct voices (leaders vs. teachers); deeper integration of quant/qual; conceptual model feels abrupt and unsupported. R2: Approach is robust but lacks methodological references. Needs precise participant numbers instead of ranges. Emirates should be presented clearly. Claims in terms of contextual/systematic, must have citations to support those claims. Descriptions of sampling strategies must cite as well as justify these in terms of stratified and/or purposive sampling. Instruments were not described in enough detail, nor were all scales cited. Interview guide was not included. Data analysis steps must cite the method, (e.g., invariance tests). Integration section needs more substantiation. R3: Why did surveys differ in time? Were they the same survey? Missing citations for many resources. Which two “additional emirates” were included? Where did “technostress buffering” appear from? It was not previously defined? The results presented only confirm the claims; no counter-data is provided.
Discussion & Conclusion <i>“Results from the regression and mediation analyses confirm that trust in leadership and psychological safety mediate the relationship between EI and readiness ...”</i> <i>“... Weekly digital use was explained by adoption intention and facilitating conditions but weakened by regulatory pressure ... This divergence underscores the value of mixed-methods approaches ...”</i> <i>“... Future studies should employ longitudinal designs to track how EI-related practices influence adoption trajectories over time ... Intervention research testing the impact of EI training on principals could determine whether improvements ... translate into measurable adoption gains ...”</i> <i>“... Training programs should therefore embed EI development alongside technical leadership skills. For policymakers, incorporating EI competencies into licensure frameworks and leadership standards would ensure ...”</i>	R1: The term “confirms” is too strong—use softer wording, such as “align with”. Only two citations used—the findings need stronger linkage to the existing literature. Practical implications should come before future research. It is suggested separating the practical and policy implications be separated into distinct sections. The implications should include practice indicators that reflect EI competencies, drawn from questionnaires and themes. R3: Discussion is too brief.

4.2. Thematic Analysis and Reviewer Convergence–Divergence

After the three peer-review reports were examined independently, two researchers noted similar patterns in the reviewers’ comments on the generated manuscript. These themes were categorized into four and then analyzed through a convergence–divergence perspective in order to find consensus and dissent across reviewers.

Theme 1—Citation Weakness and Unsupported Evidence Claims

All three reviewer reports identified weaknesses in citation support and grounding of evidence. This was most evident in statements concerning the educational context of the UAE, digitalization, emotional intelligence, and areas for further research. The result of citation integrity was thus determined to be weak; it was considered that a number of assertions were unsupported or under-cited, and were written in such a manner that had the rhetoric of a citation, creating an illusion of a base of evidence without actually including one.

Theme 2—Methodological Opacity and Limited Replicability

Reviewers concurred that the methods were not transparent enough to meet the standard for a publishable mixed-methods paper. Sampling procedures, description of instruments, scale selection, evidence of reliability and validity, statistical procedures, coding procedures, and integration techniques were either lacking or inadequately reported. These omissions rendered the paper neither verifiable nor replicable.

Theme 3—Selective Reporting and Confirmatory Alignment

Reviewers noted that the generated results were heavily confirmatory. Strong and theoretically convenient results were highlighted, and weak/null/contradictory findings were either not present or seriously underdeveloped. This gave a false sense of coherence, as the quantitative and qualitative components seemed to be telling the same predetermined story.

Theme 4—Structural Coherence with Formulaic Academic Voice

The reviewers noted the manuscript was appropriately structured and written in an academic tone. But, it was also said that the writing was formulaic, mechanical, and overconfident. This mixture of well-constructed organization, but poor evidentiary basis lent an air of believability to the manuscript that was not supported by its evidence base.

4.3. *Prompt Framing and Directional Bias Audit*

The review indicates that the prompt framing shaped the form, tone, and focus of the manuscript generated by AI. Prompts that called for coherence, integration, or theoretical consistency led the system to generate a paper in which the quantitative and qualitative results seemed to ‘fit’ together very well. This coalescence did not emerge via empirical triangulation; rather, it was generated in the same prompting sequence. The manuscript seldom reported null results or results that were unclear or that seriously contradicted other results. Instead, the simulated data and participant-style quotes were formed on the basis of the anticipated relationship between emotional intelligence, trust, psychological safety, and readiness for digital transformation. To the extent that this single-case audit can speak to this, it indicates that prompts may function as vectors that determine not only what the model outputs in text but also what kinds of evidence it pretends to generate.

4.4. *Auditability Markers Identified by Reviewers*

Across the reviewer reports, a few signals emerged that could potentially be used by editors and reviewers to detect simulated rigor within AI-mediated scholarship. None of these signs mean that a manuscript is definitely AI-written. Rather, they are cues for reviewers to ask for verification, clarification, or supporting materials. The markers identified in this investigation were:

1. Strong methodological terminology without sufficient procedural detail;
2. Sample sizes reported as ranges rather than exact figures;
3. Instruments listed without item-level description, reliability evidence, or source clarity;

4. Qualitative quotations that appear vivid but are not traceable to transcripts or participants;
5. Findings that align too neatly with hypotheses;
6. Absence of null, weak, or contradictory findings;
7. Broad contextual claims with limited citation support;
8. Generic academic phrasing that creates fluency without analytic depth;
9. Unclear connection between prompts, generated data, and final claims.

4.5. Auditability Criteria Outcomes

To address the study's focus on auditability more directly, the reviewer comments were also mapped against the auditability criteria introduced in Section 3.3. This mapping did not function as a numerical scoring exercise. Instead, it provided a structured way to identify where the generated manuscript was most and least auditable. Table 4 presents the results of this mapping. The labels in the outcome column were added by the research team after reviewing the comments against the auditability criteria described in Section 3.3. These labels are not scores. They are audit judgements used to show how visible or verifiable each element was in the generated manuscript. A criterion was marked as "weak" when reviewers pointed to major gaps, inconsistencies, unsupported claims, or evidence that could not be verified. A criterion was marked as "partial" when some evidence existed in the wider audit trail but was not clear enough in the AI-generated manuscript itself. In this sense, the assessment was based on the pattern of reviewer comments, the section-by-section mapping, and the auditability criteria used in the Methodology.

Table 4. Auditability criteria outcomes in the AI-Generated Manuscript.

Auditability Criterion	Outcome in the Generated Manuscript	Evidence from Reviewer Comments	Interpretation
Citation integrity	Weak	Reviewers identified unsupported claims, missing references, and citation-like statements without adequate evidence.	The manuscript created the appearance of an evidence base, but several claims could not be verified through cited sources.
Data generation transparency	Weak	The manuscript presented simulated statistics and qualitative quotations as if they were empirical findings.	The distinction between generated evidence and empirical evidence was not visible within the AI-generated manuscript.
Sampling consistency	Weak	Sample sizes were reported as approximate ranges, including 800–1200 participants and 50–60 schools.	Ranged sample sizes made it difficult to assess response rates, sampling adequacy, or analytic validity.
Instrument transparency	Weak	Instruments were listed but were not consistently accompanied by item details, scale structure, reliability evidence, or source clarity.	The measurement strategy appeared sophisticated but was not sufficiently auditable.
Qualitative evidence traceability	Weak	Quotations sounded authentic but were not traceable to transcripts, field notes, or real participants.	The manuscript produced synthetic qualitative voices that resembled empirical evidence without an evidence trail.
Claim–evidence alignment	Weak	Reviewers noted that several claims about context, methods, and findings were asserted rather than demonstrated.	The manuscript often relied on plausibility and coherence rather than direct support.

Table 4. *Cont.*

Auditability Criterion	Outcome in the Generated Manuscript	Evidence from Reviewer Comments	Interpretation
Selective reporting	Weak	Positive and hypothesis-confirming findings were foregrounded; null, weak, or contradictory findings were largely absent.	The manuscript generated an over-confirmatory pattern of evidence.
Prompt transparency	Partial	The study retained prompt records, but the AI-generated manuscript itself did not disclose how prompts shaped the output.	Prompt influence became visible through the audit, not through the generated manuscript's own reporting.

Note. The audit outcome was not intended as a quantitative score. It was derived through a structured matrix that linked each auditability criterion to reviewer comments, manuscript evidence, and operational definitions presented in Section 3.3. “Weak” indicates that the manuscript did not provide sufficient evidence for the criterion to be verified. “Partial” indicates that some evidence existed in the study audit trail, such as retained prompts, but was not sufficiently disclosed within the AI-generated manuscript itself.

4.6. Synthesis with Research Questions

RQ1—Scholarly Performance

In relation to RQ1, the generative system performed strongly at the level of scholarly form within this single-case audit. It generated an organized manuscript structure, plausible research questions, named instruments, statistical claims, qualitative quotations, tables, and an integrated discussion. However, the audit showed that this performance was primarily rhetorical and structural rather than evidentiary. The manuscript simulated the appearance of a mixed-methods study without providing the evidentiary transparency required for credible research.

RQ2—Answerability and the Integrity Hazard

In relation to RQ2, reviewers identified several markers of simulated rigor, including unsupported claims, missing methodological detail, inconsistent sampling information, selective reporting, and fabricated or unverifiable qualitative evidence. As can be observed in Table 4, these signs were more pronounced in the sections of citation integrity, transparency of data generation, traceability of qualitative evidence, and claim–evidence alignment. Together, these results highlight a need for peer review to focus not just on coherence and writing quality but also on the auditability of claims, data, instruments, and analyses.

RQ3—Prompt Framing and Confirmatory Alignment

With respect to RQ3, the framing effect was evident in the analysis, as it shaped the tone and orientation of the article. The produced study was biased to confirm its own assumptions, with both quantitative and qualitative findings that aligned and looked coherent but were not empirically founded. This result identifies prompt design as a key contributor to directional bias in AI-generated research.

RQ4—Auditability Standards for AI-Mediated Scholarship

With respect to RQ4, the results indicate that AI-mediated scholarship calls for more stringent auditability. These include citation checks, full disclosure of data provenance, qualitative evidence that can be traced, well-described instruments, a full description of the analytical procedures, early disclosure, and consideration of negative or inconsistent results. Lacking these protections, the superficial signs of scholarly authority may mask the lack of confirmable evidence.

Evidently, the research questions sustained a pattern of simulated rigor: the paper took the form of a plausible methodology but without the material evidentiary it would

need consider. The paper was internally consistent, but its consistency was not adequately supported by verifiable references, transparent procedures, traceable qualitative evidence, or responsible data production.

5. Discussion

In this Discussion, the results are discussed in light of the study's focal question: how can AI-generated scholarship seem credible, even if it does not rest on verifiable evidence? Because only one AI-generated paper was evaluated in the study by three expert reviewers, the results indicate rather than generalize. The purpose is not to make claims about AI-generated scholarship in general but to detect risks for auditability and markers for further exploration in a wider field of manuscripts, disciplines, and AI systems.

5.1. Generative Systems as Simulated Scholars

In this single-case audit, the results indicate that a generative system can work as a simulated scholar: it could mimic the surface appearance of academic research without necessarily producing knowledge that is empirically based. The sources supporting its claims may not be easily accessible, and the methodological means by which conclusions are drawn are frequently inadequately transparent. Therefore, the presented arguments may not be as deeply or critically interrogated as might be expected in robust scholarly investigation (Bozkurt, 2023; Rudolph et al., 2023). This reflects previous concerns that LLMs can generate polished academic prose that is not based on the evidence or conceptual reasoning that is required in scholarship (Amirjalili et al., 2024; Bender et al., 2021; Hosseini et al., 2023).

Distinction between form and evidence is crucial. The problem is not simply that the system fabricated facts, but that it fabricated a paper that looked like science. This is a danger for reviewers and readers, as the familiar structure of a research paper might trigger trust before looking at the base of evidence (Rudolph et al., 2023; Wu et al., 2023). The study therefore supports shifting the question from 'can we detect AI-generated text?' to 'can we audit for AI-mediated claims?' (Committee on Publication Ethics, n.d.; International Committee of Medical Journal Editors, n.d.; Springer Nature, n.d.).

5.2. Prompt Framing, Confirmatory Writing, and Research Integrity Risks

A key result of the audit is that prompt framing steered the narrative arc of the manuscript. When asked to generate a coherent mixed-methods manuscript based on a specific theoretical relation, the generative system produced findings that largely corroborated that relation. Instead of uncertainty, inconsistency or methodological skepticism being generated, the paper communicated a smooth convergence of hypotheses, statistical findings, qualitative quotes, and implications. This trend is consistent with concerns that LLMs may produce fluent and plausible explanations based on pattern-based reasoning rather than transparent grounded-in-evidence analyses.

The AI-generated manuscript conveyed a high degree of certainty about its hypotheses, as if they were grounded on firm empirical reality:

"Based on theory and empirical evidence, three hypotheses are proposed and tested: (H1) leaders with higher EI will be associated with greater teacher readiness for digital transformation; (H2) trust and psychological safety mediate the relationship between EI and teacher reform adoption; and (H3) regulatory pressure and cultural diversity moderate both paths."

Moreover, the reviewers asserted that competing accounts, alternative theoretical views, and disconfirming results were not sufficiently represented in the manuscript. A few core concepts necessary for the hypotheses (such as regulatory pressure and cultural

diversity) were also underdeveloped in the theoretical framework. This created a false sense of theoretical coherence, but it was not critically interrogated as would be necessary in robust scholarly work. This pattern appears to echo more general alarms that academic texts written by AI can be perfectly structured and convincing while being superficial or poorly evidenced (Amirjalili et al., 2024; Perkins, 2023).

The confirmatory nature of the manuscript was integral to the prompting process. Prompt engineering was not a neutral command line but rather a knowledge input that shaped which constructs were foregrounded, the formatting of results, and whether convergence or divergence appeared in the empirical results. This aligns with the idea proposed by White et al. (2023) that prompt engineering can improve the quality and transparency of the output, but it does not remove the need for human validation, critical thinking, and methodological responsibility. In the current study, prompts for coherence, integration, or clear mechanisms led the system to generate a paper in which the quantitative and qualitative strands seemed to converge on a common predetermined storyline.

This is a major worry for research integrity. True research can produce results that are partial, surprising, contradictory, or difficult to understand. By comparison, complexity was simply “simulated” in the generated paper. Qualitative synthetic voices were paired with synthetic quantitative trends to create the illusion of triangulation. This reinforces concerns that AI-generated outputs can create an impression of evidentiary completeness even when the underlying data, citations, and analytic procedures are not verifiable (Gao et al., 2023; Haman & Školník, 2023; Májovský et al., 2023).

Selectively reported results also came out of the generative experiment. The existence of null, weak, or even conflicting results was not acknowledged enough, while the mediator paths connecting emotional intelligence, trust, psychological safety, and readiness were assumed to be always relevant. For instance, such sentences as “these converging results taken altogether clarify the process through which EI facilitates readiness” reinforced the idea that the text was more confirmatory than informative. This is in line with the criticism directed at the ability of generative models to provide a skewed synthesis that looks balanced and fluent but limits interpretation (Wu et al., 2023; White et al., 2023).

Overall, the audit suggests that prompts can become active agents of bias in AI-mediated scholarship. They can reinforce pre-existing assumptions, reduce engagement with uncertainty, and generate the appearance of coherent evidence where no empirical grounding exists. For this reason, prompt transparency should be treated as part of research integrity. When generative AI is employed in research design, data simulation, the analysis of data, the review of literature, or writing of the manuscript, the use of the technology must be disclosed. It is not sufficient simply to state that AI was used; authors should indicate how prompts influenced the development of the structure, narrative, and claims with supporting evidence of the final output (Committee on Publication Ethics, n.d.; International Committee of Medical Journal Editors, n.d.; Springer Nature, n.d.).

5.3. Homogenization of Academic Voice and the Loss of Critical Diversity

The review also warns about the risk of a standardized academic voice. The resulting manuscript was polished, coherent, and consistent with academic norms, but reviewers noted that it often sounded formulaic and overly confident. This matters because academic writing is not only a way of reporting information, but also a place for argument, doubt, disagreement, contextualization, and intellectual stance. Should the outputs of a generative system become the standard academic prose, they have the potential to mute uncertainty, disagreement, and situated judgment. This aligns with Perkins’ (2023) claim that LLMs could level out academic writing and result in a standardization of ‘voice’ in academia. It mirrors broader concerns that the academic writing generated by AI, although well-

structured and persuasive, may be shallow, slightly less creative, and not as critically engaged as that of human writers (Amirjalili et al., 2024; Rudolph et al., 2023).

The output is also homogeneous, with recurring adoption of well-known academic keywords such as “readiness”, “trust”, “psychological safety”, “digital transformation”, and “leadership effectiveness”. These were authoritative statements but had very little dialogic engagement with alternative perspectives, situational contexts, or counter-arguments. Moreover, the article made unsupported assertions in relation to the UAE concerning cycles of reform, pressures for accountability, cultural diversity, and leadership. In short, a scholarly text that should have been well-grounded in local knowledge ended up rooted in shallow evidence.

An illustrative case was the employment of fabricated or non-existent citations to back contextual and theoretical assertions. The manuscript that emerged read:

“Research from the UAE corroborates these results:

Al Shehhi et al. (2021) argue that principals’ EI predicts critical aspects of school climate; Blaik-Hourani et al. (2020) demonstrate how the demands of accountability and cultural diversity constrain the expression of EI in schools in Abu Dhabi; and Blaik-Hourani et al. (2023) showcase both strengths and areas for development in leaders’ EI capacities against ADEK’s professional standards. Collectively, this research positions EI as a core leadership capability to lead schools through emotionally fraught changes, as opposed to a nice-to-have soft skill.”

The problem with the fragment then is not just whether individual citations are correct. The larger issue is that these citations are cited in a way that gives the appearance of having an authoritative evidence base for the particular area of inquiry under consideration. This is consistent with what Hamaň and Školník (2023) reported with regard to the ability of generative systems to produce citations, which may be inaccurate and even fabricated. The same is true of Gao et al.’s (2023) results regarding the potential for AI to produce text that sounds plausible, but is empirically unsupported.

Therefore, the study results relate to the larger issue of the credibility of knowledge. The homogenization of academic voice might limit academic argumentation and cause scholars’ writing to become less diverse and critical due to the constant repetition of similar patterns, arguments, and citations generated through AI systems. This might negatively affect the diversity of academic thought, discourse and methodology in academic writing (Perkins, 2023; Lund et al., 2023). On top of that, the study’s results further support the idea, as suggested by Stokel-Walker (2023) and Hosseini et al. (2023), that the speed of AI-powered scholarly writing should never replace human effort.

Overall, the audit suggests that a homogenized academic voice is not merely a stylistic issue. It concerns research integrity because formulaic and polished prose can make unsupported claims appear more credible, obscure gaps in evidence, and reduce the visibility of alternative interpretations. Accordingly, reviewers and editors need to be attentive not only to issues of writing but also to whether the manuscript exhibits critical diversity of perspective, contextual specificities, traceable citation practices, and a substantive engagement with opposing views.

5.4. Fabricated Evidence, Synthetic Voices, and the Problem of Verification

Among the most severe danger identified is the creation of fake or barely verifiable evidence. The AI-produced manuscript contained simulated statistics, quotes from participants, contextual assertions, and citations, all formatted as if they were real empirical reports. Nevertheless, these components involved no actual participants, data gathering, or transparent data analysis. This is indicative of broader anxieties that generative systems can make research text that looks legitimate but includes fictitious citations, unsupported

assertions, or evidence that cannot be verified (Gao et al., 2023; Haman & Školník, 2023; Májovský et al., 2023).

Reviewers were able to identify a number of claims that were not well evidenced. As an example, the automated manuscript claimed that “the strongest barriers are more organizational and human than technical,” but the citation support for this statement was insufficient. Such statements are concerning, as they sound plausible from a conceptual perspective and consistent with the context, yet they are not based on an accessible evidence base. This trend is consistent with Gao et al.’s (2023) finding that AI-generated academic text can be believable in form while lacking genuine empirical grounding.

It was also highlighted that a few of the constructs were inconsistently operationalized. Therefore, for example, a certain level of regulatory pressure was assumed and introduced as a moderator variable but was not sufficiently theorized in the conceptual model section. Cultural diversity, as a second construct, was seen as the context but was not addressed before analyzing its impact on the research. That is to say, conceptual under-specification is purpose-serving in that it allows for disclosure of theoretical complexity in the manuscript without necessarily requiring the disclosure of complex thought. In that way, it ties to concerns about LLMs inventing entire plausible theories of pattern (Hosseini et al., 2023; Rudolph et al., 2023).

The most significant verification problem concerned the generation of qualitative evidence. The manuscript included participant-style quotations such as:

“I decided to try the new e-learning platform because my principal encouraged us personally.”

“I openly admit when I struggle with new tools myself . . . it shows them it’s okay not to be perfect.”

These quotes were constructed to sound like real voices. It seemed they added qualitative richness and served as a triangulation of the quantitative results. Yet, since the research was simulated, these voices did not correspond to any real interview, group discussion, transcript, or person. The use of synthetic qualitative quotes is particularly dangerous, as they create an emotional sense of authenticity. An evocative quote can convince one that he has heard the voice of the participant even though there is no participant at all.

Similarly, the correspondence between the qualitative quotations and the quantitative results contributed to a sense of triangulation. While the paper indicated that the survey results and the qualitative findings pointed to the same interpretations, in mixed-methods research, the integration of evidence relies on clear connections between separate strands. In the case of the generated paper, convergence occurred in the same generative process, resulting in something that might be termed synthetic credibility: evidence that looks real and credible but lacks independent verification.

The paper also created some minor tensions, such as the idea that regulatory pressure was only having a small quantitative impact but was more significant from a qualitative perspective. On the surface, this might seem like adding realism and complexity to the analysis. But in this case, the complexity stemmed from the fact that neither the qualitative nor quantitative data existed and had been created for the purposes of the paper. This is not to say that there can be no contradictions or discrepancies in real research papers; the problem lies in the fact that none existed here.

These patterns demonstrate the need for stronger verification procedures. There needs to be an audit trail for qualitative data from transcripts, anonymous quotes, field notes, and so on. Transparency is needed regarding quantitative information on how data were generated, a dictionary of the variables used, the coding and analysis procedures, and the disclosure of any results that contradict the hypothesis. Instead, they should investigate

and determine the level of risk associated with harmful content generation, as far as biased and one-sided narratives are concerned, which appear legitimate while in turn confirming hypotheses rather than contributing to trusted knowledge (Hosseini et al., 2023; White et al., 2023).

In summary, the review reveals that falsified evidence is not just an accuracy problem. It is a matter of scientific reliability. When synthetic statistics, synthetic voices, or weakly verified citations are wrapped in the familiar language of empirical research, style and presentation can confuse readers, reviewers, and editors into mistaking plausibility for validity. Consequently, AI-enabled scholarship needs more rigorous standards of verification such as checking citations, data transparency, rapid disclosure, traceable qualitative evidence, and an auditable chain of relationships between claims, methods, and evidence.

5.5. Methodological Transparency and the Limits of Surface Rigor

Previous audit studies suggest that LLMs can perform relatively well on structured assessments, such as professional examinations, while still struggling with tasks that require contextual judgment, evidence verification, and methodological transparency (Bommarito & Katz, 2022; Gilson et al., 2023; Kung et al., 2023). The present audit illustrates this distinction in the context of a generated research manuscript.

The section on methodology unambiguously declared an “explanatory sequential mixed-methods design” using an integrative approach through “joint displays and meta-inferences.” Although the nomenclature replicated gold standard references (Creswell & Plano Clark, 2011), reviewers highlighted the fact that no methodological sources or examples of research were cited. There was a claim of an integrative approach; however, evidence for this was neither presented nor validated, thus proving the assertion by Hosseini et al. (2023) that while generative systems exhibit linguistic expertise of a scholar, veracity cannot be determined.

Such ranges as “N = 800–1200 teachers; 50–60 schools” were unfeasible for assessing whether the sampling was sufficient or not, and whether there was a high rate of respondents. In terms of measures, several scales are mentioned (WLEIS, Edmondson, ORIC, UTAUT, technostress, and regulatory pressure), but no information is provided about the scales and the creation of a single survey by combining all items. There were differences in the lengths of surveys between leaders and teachers for unclear reasons.

In the analysis section, the use of sophisticated terminologies like “multilevel SEM,” “measurement invariance,” and “bootstrapped indirect effects” was observed; however, no mention of fit statistics, different types of SEMs, or the invariance process was noted. Similarly, although Braun and Clarke’s thematic analysis, or the qualitative phase, is yet to present a theme organization or a coding scheme, some quotations have been offered by the authors to support the quantitative findings.

Ultimately, the “Timeline and Deliverables” section created an unproductive, vague, and generic boundary between a research proposal and a research article. As Perkins (2023) explains, generative writing systems produce a carefully constructed document that undermines the very foundations of scholarly research.

Taken together, these methodological indicators show how generative systems can imitate rigor without reproducing the conditions that make rigor meaningful. Terms such as “multilevel SEM,” “measurement invariance,” “thematic analysis,” and “joint displays” carry scholarly authority, but they require procedural detail to be credible. Without information about sampling, instruments, data structure, coding, fit indices, integration procedures, and contradictory findings, methodological language becomes decorative rather than evidentiary.

5.6. From Disclosure to Auditability

Current publishing guidance increasingly requires authors to disclose the use of AI-assisted technologies, but disclosure is only the first step. A paper can reveal that AI was used but still offer citations that cannot be checked, synthetic data, or conclusions that are lightly supported. At the same time, you can have a paper with no AI involvement that is not sufficiently transparent. Integrity of research, therefore, needs a more expansive auditability framework.

Auditability is the capacity to trace claims in a manuscript to the evidence and processes of inquiry that are open to examination. In AI-mediated scholarship, this means determining not just which tools were used but also how they were prompted, what outputs were generated, how those outputs were vetted, and which portions of the final manuscript rely on AI assistance. It brings ordinary, but essential, good-practice aspects of research integrity into play: verification of citations, availability of data, transparency of instrument, reproducibility of analysis, and presenting a balanced report of results.

5.7. Implications for Editors, Reviewers, and Institutions

These results are directly relevant to academic publishing. Editor and reviewers should not take polished academic style as a mark of quality. They should instead ask whether the paper has an auditable trail that leads from claims back to evidence. This is especially important for papers reporting complicated procedures, large datasets, qualitative quotations, or particularly coherent mixed-methods integration. For institutions, the results indicate that AI literacy should be combined with research integrity education. Researchers must understand that generative AI can assist in writing and organization, but it cannot replace empirical evidence, ethical data collection, critical interpretation, or authorial responsibility. Training should therefore address not only how to use AI tools but also how to verify, document, and responsibly disclose their use.

5.8. AI-Specific Risks and General Research Quality Issues

A key counterinterpretation should be considered: many of the weaknesses identified in the produced manuscript are not specific to research generated by AI. Unsupported claims, vague methods, selective reporting, weak integration, and limited engagement with alternative interpretations can also appear in human-authored manuscripts. The presence of these weaknesses, therefore, does not prove by itself that a manuscript was AI generated.

The contribution of this study is not to claim that these problems are exclusive to generative AI. Instead, it shows how generative systems can reproduce many of these weaknesses quickly, fluently, and simultaneously while also producing synthetic citations, synthetic participant voices, simulated datasets, and prompt-shaped coherence. The AI-specific risk lies in the combination of scale, fluency, fabrication, and prompt-induced alignment. These features may make weak or fabricated research appear more credible than it is (Table 5).

This distinction is important for responsible peer review. The auditability markers proposed in this article should not be used to accuse authors of AI misuse. They should be used as prompts for verification. A manuscript exhibiting these markers might require more supportive data, clearer methodology, greater support from citations, or more transparent reporting, whether or not AI was utilized.

Table 5. Distinguishing general research- quality issues from AI-specific amplifications.

Issue	Also Found in Human-Authored Research?	AI-Specific Amplification	Audit Response
Unsupported claims	Yes	Claims may be generated fluently without source grounding.	Verify claim-source alignment.
Methodological ambiguity	Yes	Methodological terminology may be generated without procedural depth.	Request detailed protocols, instruments, and analysis procedures.
Selective reporting	Yes	Generated outputs may align too neatly with the prompt or expected conclusion.	Ask for null, weak, contradictory, or unexpected findings.
Fabricated citations	Possible but less common	Generative systems may produce plausible but non-existent or misrepresented citations.	Check source existence and claim–source relevance.
Synthetic quotations	Rare and unethical if undisclosed	Generative systems can produce realistic participant-style voices without participants.	Require transcripts, anonymized excerpts, or a qualitative audit trail.
Prompt-induced coherence	Primarily AI-specific	Prompt wording can shape constructs, evidence, and conclusions.	Require AI-use and prompt disclosures where relevant.

6. Recommendations: An Audit Checklist for AI-Mediated Scholarship

The results indicate that journals, reviewers, and institutions require practical auditing processes for assessing AI-assisted scholarship. Classical features of quality of a manuscript, such as fluency, organization, use of methodological language, and coherence, should still be regarded as important, but these are no longer sufficient by themselves. Emerging from this single-case audit was a manuscript that was scholarly in form, yet it was weakly supported by citation, simulated evidence, contained variable methodological detail, and was strongly confirmatory in reporting. Therefore, audibility should be considered as a further research integrity issue: authors should be able to demonstrate that the claims, citations, data, methods, and interpretive conclusions of a manuscript can be linked to verifiable evidence and transparent procedures.

The following checklist is offered as a provisional aid for reviewers and editors, with the aim of highlighting aspects where further clarification, documentation, or confirmation might be warranted. It is not meant to accuse authors of misuse of AI. It is to promote responsible consideration and review of manuscripts for which the use of generative AI is disclosed or suspected, as well as manuscripts in which other transparency and evidentiary support issues arise.

The checklist is evidence-informed but has not been formally validated. It was informed by four sources: audit findings reported in this paper; convergence of reviewer comments, prior work on AI weaknesses in scholarly outputs, and existing guidance in the publication ethics space. It has not been validated through completion of a Delphi process, analysis of inter-rater reliability, or application to multiple manuscripts and disciplines. For this reason, the checklist should be understood as a preliminary framework that requires further validation and refinement (Table 6).

Table 6. Audit checklist for evaluating AI-mediated scholarship.

Audit Area	Core Reviewer/Editor Question	Additional Verification Prompts	Evidence or Documentation to Request
Citation integrity	Do all cited sources exist, and do they support the specific claims made in the manuscript?	Are any citations fabricated, misattributed, outdated, or irrelevant? Do cited sources support the exact claim, not merely the general topic? Are direct quotations and page references accurate where applicable?	Reference verification sample; DOI/cross-ref check; corrected reference list; page numbers for direct quotations.
Data authenticity	Is there evidence that the dataset is real, ethically collected, and available for verification where appropriate?	Does the manuscript clearly distinguish real, synthetic, simulated, or AI-generated data? Is the data source described clearly? Are ethical approvals, consent procedures, or data access limits explained?	Ethics approval information; anonymized dataset; data dictionary; explanation where data cannot be shared.
Qualitative evidence	Are participant quotations traceable to transcripts, anonymized excerpts, field notes, or a secure audit trail?	Do quotations correspond to real participants or are they synthetic examples? Are participant groups clearly identified? Are quotations linked to themes, codes, or transcript evidence?	Anonymized transcript excerpts; field notes; coding framework; statement on whether quotations are real or synthetic.
Methodological consistency	Are sample sizes, instruments, procedures, and analyses reported consistently across the manuscript?	Are sample sizes exact rather than approximate? Are participant numbers consistent across Abstract, Methods, Results, tables, and figures? Are the reported analyses appropriate for the stated design?	Revised Methods section; participant flow information; consistency check across manuscript sections.
Instrument transparency	Are survey items, interview guides, codebooks, scales, and reliability evidence clearly described or provided as supplementary material?	Are all instruments named and cited? Are item numbers, response scales, and example items reported? Is reliability or validity evidence provided for each scale?	Survey instrument; interview / focus-group guide; scale citations; reliability / validity evidence; supplementary materials.
Analytic transparency	Are statistical tests, model specifications, coding procedures, and integration strategies explained in enough detail to permit scrutiny?	Are model-fit indices, estimation methods, and assumptions reported where relevant? Are coding procedures and theme development described? Is mixed-methods integration demonstrated rather than only claimed?	Statistical syntax or analysis plan; codebook; joint displays; integration matrix; explanation of analytic decisions.
Prompt transparency	If AI was used, are the model, purpose of use, prompt strategy, and degree of human oversight disclosed?	Which AI tool was used? For which tasks? Were prompts used for design, data generation, analysis, interpretation, or writing? How were outputs checked by human authors?	AI-use declaration; prompt archive or prompt summary; model/version information where available; description of human verification.
Selective reporting	Are null, weak, contradictory, or unexpected findings reported, or does the manuscript only present confirmatory evidence?	Does the manuscript report non-significant findings? Are tensions between quantitative and qualitative findings acknowledged? Are limitations and alternative interpretations discussed?	Full results tables; analysis plan; explanation of excluded findings; discussion of null or contradictory results.
Supplementary materials	Are raw data, data dictionaries, coding frameworks, prompts, or analytic files available where ethically and legally possible?	Are the supplementary materials sufficient to check claims? If materials cannot be shared, is the reason clearly explained? Are access procedures provided?	Data repository link; supplementary files; restricted access statement; prompt or output archive where appropriate.
Scholarly voice and critical judgment	Does the manuscript show critical engagement, uncertainty, and contextual judgment, or only polished but generic academic phrasing?	Does the manuscript engage with alternative interpretations? Are limitations meaningfully discussed? Does the writing show disciplinary judgment rather than formulaic coherence?	Revised discussion; stronger engagement with counter-evidence; clearer limitations; contextualized interpretation.

These criteria should not be treated as evidence of misconduct. They identify areas where a manuscript may require additional explanation, verification, or documentation. A poor answer to one item on the checklist does not demonstrate that a manuscript has been written by AI or that an author has acted unethically. Rather, the checklist supports editors and reviewers in determining at a glance where in a scholarly work claims need to be better supported with evidence. Its role is to shield academic inquiry from the danger that plausible scholarly form is separated from confirmable evidence. Disclosure

is necessary but not sufficient. The use of generative AI in research design, literature searches, instrument development, data generation, data analysis, coding, translation, interpretation, or writing of the manuscript should be disclosed by the authors. However, disclosure must be accompanied by human review and verification. If AI has affected the structure, trajectory, or evidentiary assertions of a manuscript, authors should also describe the prompt strategy, the type of generated outputs, and the procedures to verify accuracy, source trustworthiness, and methodologic coherence.

Verification of the evidence is particularly warranted. Participant quotations in qualitative work should be traceable to transcripts that are ethically stored, anonymized excerpts, field notes, or some other form of credible audit trail. In quantitative research, authors should make raw data, data dictionaries, and analytic code available or provide a valid explanation for why these information cannot be shared. If synthetic or simulated data are employed for valid reasons, they should be clearly marked as synthetic and not be represented as results from empirical studies.

Reviewers should also be vigilant for excessive coherence. Real empirical results are often partial, complex, mixed, or hard to interpret. Papers reporting only positive results, a tightly convergent qualitative and quantitative evidence, or results that perfectly confirm the hypotheses merit scrutiny, particularly if methodological detail is scant. Author must be urged to report also null, weak, inconsistent or unexpected results where these have been obtained. Preregistration, registered reports, analysis plans, and transparent reporting of excluded or non-confirmatory results may mitigate this risk.

Journals and institutions should also provide incentives for reviewer training in simulated rigor. Reviewers could also be provided with practical advice on how to identify concerns with audibility, such as inconsistent reporting of samples, ambiguous descriptions of instruments, non-verifiable or fraudulent citations, fictitious participant quotes, over-reliance on a formulaic academic voice, prompt-shaped coherence, and use of methodological terms that lacks any procedural content.

Therefore, generative AI may be able to facilitate the planning, writing and editing of scholarly output, but it is not a substitute for human rationality, empirical investigation, methodological rigor, or ethical responsibility. The aim should not be to disallow AI-enabled research, but to ensure that any AI-mediated input is transparent, verifiable, and accountable to the norms of scholarship. The suggested checklist provides an initial step in that direction by helping reviewers and editors to ask not just if a manuscript is well-written but if its claims can be traced, checked, and justified.

7. Limitations and Directions for Future Research

This investigation needs to be viewed in the context of its design and objectives. It takes the form of a single-case study analysis, focusing on one AI-generated manuscript and a small number of expert reviewers. This allows for a concentrated investigation of audibility and simulated rigor, but does not make claims to be comparative or generalizable across all generative systems or contexts. The audit checklist proposed in Section 6 should also be understood as preliminary and evidence-informed rather than formally validated; future research should test its clarity, usability, inter-reviewer consistency, and diagnostic value across larger and more diverse manuscript samples.

Employing the simulation of the research artifact is a conscious methodological decision that enables issues of verification, transparency, and evidentiary traceability to be explored in depth. Yet it fails to capture the full intricacy of AI-enabled research workflows, where human decision-making, iteration, and domain knowledge converge with generative tools.

The analysis is based on the interpretation of expert reviewers, consistent with well-established practices of scholarly evaluation. Future work could extend this approach through additional systematic, comparative, or cross-model analyses to investigate variation in generative performance and auditability risks.

Further investigation might push this line of enquiry by exploring other types of generative systems, prompt patterns and disciplinarity. Furthermore, it may study the practical implementation of auditability frameworks in research and publishing, possibly including the development of tools, standards, and institutional practices that make transparent and verifiable AI-assisted scholarship.

8. Conclusions

This study explored the performance of a generative AI system when accepted as a simulated scholar in an exploratory split-paper peer-review audit. The analysis was based on a single AI-written mixed-methods research manuscript and three independent expert reviews. In this single-case audit, the produced manuscript effectively reproduced multiple formal and stylistic conventions of academic writing, such as research questions, hypotheses, methodological language, statistical statements, qualitative quotations, tables, and discussion. Nevertheless, the audit also found that these formal attributes were not accompanied by the evidentiary support, methodological transparency, citation integrity, and critical judgment necessary to make scholarship credible.

The manuscript appeared to be credible as it had the typical structure of a research paper. But the investigation found that many of its assertions, data points, quotes, and methodological steps were simulated, not supported by empirical evidence. The trustworthiness of the manuscript thus hinged on structure, flow, and internal coherence rather than on verifiable evidence. This distinction between a plausible form of scholarly and an auditable evidence of the type is the key focus of the paper.

The research reveals three dominant threats in the investigated case. To begin with, surface plausibility can create the impression of rigor and hide weak, unsupported, or fabricated evidence. Second, prompt framing may encourage confirmatory alignment of outputs, resulting in outputs and interpretation that are consistent with the assumptions contained in the prompt. Third, synthetic evidence, such as made-up statistics and quotations mimicking participant responses, can obscure the line between empirical and simulated research. These are risks to be considered in the context of, and not as the defining features of, all AI-mediated scholarship, but rather the concerns around auditability that must be handled with care in research and peer review.

The main novelty of this paper is turning the focus from detection to auditability. Instead of just asking whether a given manuscript was written by AI, the study enquires whether its assertions, citations, data, methods, qualitative evidence, and interpretive conclusions can be followed and checked. The split-paper audit protocol is one means to render this process visible, situating the AI-generated manuscript next to expert reviewer assessments. The reporting guidelines for the systematic review proposed in this article is a preliminary evidence-based tool and not a validated instrument. It provides a basis for potential further testing, refinement, and application within other disciplines, types of manuscript, groups of reviewers, and generative AI systems.

Generative AI may aid with some components of scholarly labor, such as writing, organizing, editing, translating and initial synthesis. But it cannot take responsibility for ethics, validate evidence, justify methodological choices, or substitute human scholarly judgment. The recommended response is thus not to dismiss the AI-enabled scholarship but to demand higher levels of transparency, verifiability, and accountability. Yet, as generative systems are producing ever more authentic scholarly form, the key challenge for scholars,

reviewers, editors, and institutions will be to keep scholarly claims tied to real evidence, transparent methods, and responsible human oversight.

Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

In the process of writing this paper, the generative AI model ChatGPT was employed to generate a hypothetical mixed-methods research manuscript as the main piece of analysis. The manuscript itself contains synthesized data, fictional quotations from participants, methods of research, and an AI-generated academic text. The generative AI model has also been applied to assist in the language editing and structuring processes during the development of the manuscript. Ultimately, all the AI-generated information has been critically analyzed and assessed by human authors, who are thus fully responsible for the final product.

Author Contributions: Conceptualization, R.H.; methodology, R.H.; software, Y.M.A.; validation, R.H. and P.F.; formal analysis, E.S.E., F.E.Z. and R.H.; investigation, R.H., Y.M.A., P.F. and A.M.H.; resources, Y.M.A.; data curation, R.H., P.F. and Y.M.A.; writing—original draft preparation, R.H., P.F. and Y.M.A.; writing—review and editing, E.S.E., F.E.Z. and A.M.H.; visualization, P.F. and R.H.; supervision, R.H.; project administration, R.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Alnaim, N. (2024). Generative AI: A case study of ChatGPT's impact on university students' learning practices. *Research Square*. [CrossRef] [PubMed]
- Al Shehhi, M., Alzouebi, K., & Ankit, A. (2021). An examination of the emotional intelligence of school principals and the impact on school climate in public schools in the United Arab Emirates. *Journal of Applied Research in Higher Education*, 13(5), 1269–1289. [CrossRef]
- Amirjalili, F., Neysani, M., & Nikbakht, A. (2024). Exploring the boundaries of authorship: A comparative analysis of AI-generated text and human academic writing in English literature. *Frontiers in Education*, 9, 1347421. [CrossRef]
- Bender, E., McMillan-Major, A., Shmitchell, S., & Gebru, T. (2021). On the dangers of stochastic parrots: Can language models be too big? In *FAccT'21: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623). ACM. [CrossRef]
- Blaik-Hourani, R., Litz, D., Ali, N., Azaza, M., & Parkman, S. (2023). Abu Dhabi school leaders' emotional intelligence: Dimensions, deficits, and capacity building for school improvement. *Educational Research for Policy and Practice*, 22, 259–282. [CrossRef]
- Blaik-Hourani, R., Litz, D., & Parkman, S. (2020). Emotional intelligence and school leaders: Evidence from Abu Dhabi. *Educational Management Administration & Leadership*, 49(3), 174114322091355. [CrossRef]
- Bommarito, M. J., II, & Katz, D. M. (2022). GPT takes the bar exam. *arXiv*. [CrossRef]
- Bozkurt, A. (2023). Generative artificial intelligence (AI) powered conversational educational agents: The inevitable paradigm shift. *Asian Journal of Distance Education*, 18(1), 198–204. [CrossRef]
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. [CrossRef]
- Chen, J., Zhuo, Z., & Lin, J. (2023). Does ChatGPT play a double-edged sword role in the field of higher education? An in-depth exploration of the factors affecting student performance. *Sustainability*, 15(24), 16928. [CrossRef]
- Committee on Publication Ethics. (n.d.). *Artificial intelligence*. COPE. Available online: <https://publicationethics.org/cope-focus/artificial-intelligence> (accessed on 10 June 2026).
- Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research*. Sage Publications.

- De Silva, A., Wijekoon, J., Liyanarachchi, R., Panchendrarajan, R., & Rajapaksha, W. (2024, March 4–28). *AI insights: A case study on utilising ChatGPT intelligence for research paper analysis*. 14th International Workshop on Bibliometric-enhanced Information Retrieval, Glasgow, Scotland. Available online: <https://arxiv.org/pdf/2403.03293> (accessed on 10 June 2026).
- Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44(2), 350–383. [CrossRef]
- Elsevier. (n.d.). *Generative AI policies for journals*. Available online: <https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals> (accessed on 10 June 2026).
- Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine*, 6, 75. [CrossRef] [PubMed]
- Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How does ChatGPT perform on the United States Medical Licensing Examination? The implications of LLMs for medical education and knowledge assessment. *JMIR Medical Education*, 9, e45312. [CrossRef] [PubMed]
- Haman, M., & Školník, M. (2023). Using ChatGPT to conduct a literature review. *Accountability in Research*, 31(8), 1244–1246. [CrossRef] [PubMed]
- Hosseini, M., Rasmussen, L. M., & Resnik, D. B. (2023). Using AI to write scholarly publications. *Accountability in Research*, 31(7), 715–723. [CrossRef] [PubMed]
- International Committee of Medical Journal Editors. (n.d.). *Use of AI by authors*. Available online: <https://www.icmje.org/recommendations/browse/artificial-intelligence/ai-use-by-authors.html> (accessed on 10 June 2026).
- Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health*, 2(2), e0000198. [CrossRef]
- Liu, M., Wang, T., Song, X., Jing, B., Yang, J., Song, J., Chen, K., Chen, S., Li, M., Niu, Q., Liu, J., Peng, B., Zhang, S., Xu, J., Pan, X., Wang, J., Feng, P., Zhang, L., Zhong, Y., & Wen, Y. (2025). *Ethics and social implications of large language models*. Available online: <https://ssrn.com/abstract=5207555> (accessed on 10 June 2026).
- Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74(5), 570–581. [CrossRef]
- Maistry, S. M., & Singh, U. G. (2025). Faculty perspectives on the role of ChatGPT-4.0 in higher education assessments. *Journal of Education*, 98, 86–102. [CrossRef]
- Májovský, M., Černý, M., Kasal, M., Komarc, M., & Netuka, D. (2023). Artificial Intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *Journal of Medical Internet Research*, 25, e46924. [CrossRef] [PubMed]
- Perkins, M. (2023). Academic integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice*, 20(2), 1–24. [CrossRef]
- Pietarinen, A.-V., & Shi, S. (2026). Manipulation and deception in generative AI-mediated education: Preserving epistemic agency, critical thinking, and creativity. *Postdigital Science and Education*. [CrossRef]
- Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching*, 6(1), 342–363. [CrossRef]
- Sahebi, S., & Formosa, P. (2025). The AI-mediated communication dilemma: Epistemic trust, social media, and the challenge of generative artificial intelligence. *Synthese*, 205, 128. [CrossRef]
- Springer Nature. (n.d.). *Artificial intelligence (AI)*. Nature Portfolio. Available online: <https://www.nature.com/nature-portfolio/editorial-policies/ai> (accessed on 20 June 2026).
- Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: Many scientists disapprove. *Nature*, 613(7945), 620–621. [CrossRef]
- Tarafdar, M., Tu, Q., Ragu-Nathan, B. S., & Ragu-Nathan, T. S. (2007). The impact of technostress on role stress and productivity. *Journal of Management Information Systems*, 24(1), 301–328. [CrossRef]
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425–478. [CrossRef]
- Weiner, B. J. (2009). A theory of organizational readiness for change. *Implementation Science*, 4(1), 67. [CrossRef] [PubMed]
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv*. [CrossRef]
- Wong, C.-S., & Law, K. S. (2002). Wong and law emotional intelligence scale. In *APA PsycTESTS dataset*. American Psychological Association. [CrossRef]

- Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D. F., & Chao, L. S. (2023). A survey on LLM-generated text detection: Necessity, methods, and future directions. *arXiv*. [\[CrossRef\]](#)
- Zimmermann, R., Staab, M., Nasser, M., & Brandtner, P. (2024). Leveraging large language models for literature review tasks—A case study using ChatGPT. In T. Guarda, F. Portela, & J. M. Diaz-Nafria (Eds.), *Advanced research in technologies, information, innovation and sustainability. ARTIIS 2023. Communications in computer and information science* (Vol. 1935). Springer. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.