Two Last Human Sections? A Case Study of Largely AI-Generated Paper Writing in Empirical XR Research

Anusha Devanga[*] University of Central Florida
Gerd Bruder[†] University of Central Florida
Daniel Zielasko[‡] Technical University of Denmark
Alexander Giovannelli[§] University of Central Florida
Zubin Choudhary[¶] University of Central Florida
Hiroshi Furuya[‖] University of Central Florida
Matthew Gottsacker[**] University of Central Florida
Robert W. Lindeman[††] University of Canterbury
Gregory Welch[‡‡] University of Central Florida

ABSTRACT

Generative Artificial Intelligence (AI) is rapidly transforming scientific research, particularly the analytical activities surrounding empirical studies. While experiment design, data collection, and statistical analysis remain largely human-driven in extended reality research, the introduction, related work, discussion, conclusion, abstract, and title of a paper are increasingly susceptible to AI generation. In this paper, we investigate through a practical example how an AI can generate analytical framing of a scientific paper when provided only with human-authored empirical sections. Starting from an accepted but yet-to-be-published ISMAR 2026 paper, we retained only the experiment and results sections and iteratively used a state-of-the-art AI system to generate the remainder of the manuscript. Comparing the resulting AI-generated paper with the original human-authored version, we find that the AI produces a coherent, persuasive, and publication-ready narrative, but one that systematically reframes the work around the most prominent findings rather than the researchers' original motivations. Our results show that fully AI-generated analytical framing of empirical research is already feasible, raising questions about scientific authorship, researcher intent, publication incentives, and the future role of human judgment in scientific communication.

Index Terms: Generative AI, LLMs, Extended Reality.

1 INTRODUCTION

Artificial intelligence (AI) is rapidly becoming embedded throughout the scientific research process [20, 26]. What began as assistance for spelling and grammar through large language models (LLMs) has progressed to literature discovery, synthesis of prior work, statistical analysis, code generation, interpretation of results, manuscript drafting, and even peer review [8, 23]. Across many scientific fields, researchers increasingly interact with AI systems not merely as *tools* for communication, but as *collaborators* in the production of scholarly artifacts [20]. The question is no longer *whether* AI can assist scientific work, but *which aspects* of scientific work remain fundamentally human and which are becoming delegable to machines [4, 19].

[*]e-mail: anusha.devanga@ucf.edu
[†]e-mail: bruder@ucf.edu
[‡]e-mail: danzi@dtu.dk
[§]e-mail: gio@ucf.edu
[¶]e-mail: zubin.choudhary@ucf.edu
[‖]e-mail: furuya@ucf.edu
[**]e-mail: mattg@ucf.edu
[††]e-mail: rob.lindeman@canterbury.ac.nz
[‡‡]e-mail: welch@ucf.edu

This distinction is particularly important in empirical disciplines such as human-computer interaction (HCI) and extended reality (XR) [12]. Not all scientific contributions are of the same nature:

- *Analytical Contributions* involve synthesizing existing XR literature, motivating research questions, explaining methodologies, interpreting results, discussing limitations, and positioning findings within the broader scientific context. These tasks rely heavily on reasoning, communication, and the ability to organize knowledge, making them well suited for AI.
- *Empirical Contributions* involve the collection of original data, the execution of experiments, interacting with participants and XR technologies, measuring phenomena, and observations about the world.

Recent advances in generative AI have demonstrated remarkable capabilities with respect to *analytical* work [10]. LLMs can already retrieve and summarize literature [17], generate related work sections [11, 16], construct persuasive scientific narratives [21], formulate limitations [24], propose future research directions, produce publication-ready prose [1], and support reviewing [5]. In many cases, the quality of this output is sufficiently high that readers cannot determine whether it was written by a human or generated by a machine [27]. By contrast, *empirical* contributions remain comparatively resistant to automation. While AI can assist experiment design or data collection [14], the generation of new empirical evidence still largely depends on human researchers, physical systems [18], human participants, and interactions with the real world. In disciplines such as XR, where findings often emerge from complex interactions between users, devices, environments, tasks, and social contexts, empirical research remains rooted in observations that cannot simply be inferred from existing text corpora [13].

This apparent division has created an emerging assumption within parts of the XR research community, namely that empirical research may remain a fundamentally human activity even as analytical scientific writing becomes increasingly automated. Yet this distinction creates a subtle and potentially important tension. Scientific findings are not communicated to the community as raw empirical observations alone. Empirical results require interpretation. They must be situated within prior work, connected to broader research questions, framed as contributions, and translated into the narrative structure expected by publication venues [22]. Traditionally, these analytical activities have been viewed as essential intellectual contributions of the authors. Increasingly, however, they are becoming activities that authors delegate partially or entirely to AI [7]. Indeed, many contemporary empirical papers already exhibit characteristics consistent with AI-assisted scientific writing [6]. As generative AI becomes more capable, it becomes increasingly difficult to determine which portions of a publication reflect the authors' original analytical framing and which represent an AI-generated interpretation and presentation built around the empirical results [3, 15].

This raises an important **research question** for the future of empirical science in XR: If researchers provide only the empirical foundation of a study—its methods, procedures, data, and results—what kind of paper would emerge if we let AI generate the remainder of the paper [25]? Would it resemble a paper written by the researchers? Would it identify the same scientific contribution? Would it frame the same research questions and implications? Or would it construct an alternative narrative shaped primarily by the statistical outcomes of the study rather than by the researchers' original intentions?

In this paper, we put this possibility to a direct test. Starting from an accepted, but yet-to-be-published ISMAR 2026 paper, we isolate the empirical components of the manuscript and iteratively task an AI system with generating all remaining analytical parts, including the related work, discussion, introduction, conclusion, abstract, and title. The resulting manuscript, therefore, represents a paper whose empirical foundation originates entirely from human researchers, while its analytical framing is generated entirely by AI.

Our goal is not to evaluate whether AI can produce *grammatically correct* scientific prose; that capability is already well established [10]. Instead, we investigate whether AI can independently construct the *analytical narrative* that transforms empirical findings into a publishable scientific contribution. In doing so, we explore broader questions regarding authorship, scientific framing, researcher intent, and the future role of humans in empirical science.

This paper makes two main contributions:

- We present a case study in which all analytical sections of an accepted empirical research paper are generated by AI using only the experiment and results sections as input.
- We analyze the similarities and differences between the original human-authored paper and the AI-framed paper, discussing the implications for scientific communication and the future of empirical scholarship in the age of AI.

The remainder of this paper is structured as follows. Section 2 outlines the methodology we used for our approach and the analysis we performed. Section 3 describes the results we observed and discusses our findings. Section 4 concludes our paper.

2 Method

To investigate whether the analytical portions of an empirical research paper can be generated entirely by AI, we conducted a controlled paper reconstruction study. Our starting point was a full paper recently accepted at the IEEE International Symposium on Mixed and Augmented Reality (ISMAR) 2026 [2]. Importantly, the version used in our study was the accepted manuscript rather than the final camera-ready version. As is typical in the publication process, additional revisions and editorial refinements were made before final publication. Nevertheless, the paper represented a typical empirical XR publication, containing both empirical contributions (experiment design, methodology, and statistical results) and analytical contributions (motivation, literature framing, interpretation, implications, and positioning within the broader research landscape).

Rather than simply evaluating whether AI can generate isolated manuscript components, our goal was to determine whether AI could reconstruct an entire publication-quality scientific narrative from empirical evidence alone. The AI was not given the original research question of the authors for the generation of the narrative.

2.1 Analytical and Empirical Components

We distinguish between primarily empirical and analytical components of such an XR paper. The primarily empirical parts of the paper consist of the sections that describe how evidence was generated (*Experiment*) and what evidence was obtained (*Results*). These sections typically contain the study design, apparatus, materials, procedures, participant information, hypotheses, dependent measures, descriptive statistics, inferential statistics, figures, and tables.

The primarily analytical parts of the paper consist of the front matter (*Title, Abstract*), followed by sections that explain why the work was conducted (*Introduction*), how it relates to prior literature (*Related Work*), what the findings mean/why they matter (*Discussion*), and the main "take-aways" from the paper (*Conclusion*).

For the purpose of this study, we treated the first category as human-generated empirical work and the second category as analytical narrative generated by either humans or AI.

2.2 Creation of the Basic Manuscript

As shown in Figure 1, to isolate the empirical contribution, we removed all analytical sections from the accepted paper. The resulting manuscript retained the original figures and tables, the complete *Experiment* section, the complete *Results* section, and section headings for all remaining manuscript sections. The document, therefore, resembled a partially completed conference paper in PDF format using the official ISMAR 2026 conference template. It contained sufficient information to understand the study design, collected data, and statistical outcomes, but no information regarding the authors' intended framing, literature positioning, interpretation, or claimed contributions. This basic empirical manuscript served as the sole source material provided to the AI with no top-down guidance on the research question that was investigated by the authors. The paper was generated in the following order: *Related Work*, *Discussion*, *Introduction*, *Conclusion*, *Abstract*, and finally the *Title*. We chose this sequence so that the agent could build context from both the sections it had been provided and those it had already generated. As a result, each section could provide relevant context for the next, which we anticipated would improve the overall coherence and quality of the final paper.

2.3 Iterative AI Generative Process

The generative AI system used for this study was Microsoft Enterprise Copilot powered by Claude Opus 4.7 with a knowledge cutoff of January 2026. This AI does not learn from the submitted document. Prior to the experiment, Copilot was given custom instructions to ensure the responses were appropriate (Figure 2).

> **Prompt.** You are an experienced research assistant tasked with writing a section for a highly rigorous scientific publication for the prestigious International Symposium on Mixed and Augmented Reality (ISMAR), a top scientific venue for highly impactful research contributions on virtual and augmented reality. The expectation is to write a paper that will receive the highest review scores at this conference through the highest form of scientific reasoning. The publication must follow the highest standards for scientific language and intellectual contributions. The language of the paper must match the scientific terminology of the wider literature corpus of papers published at the ISMAR conference. Every argument, claim, and interpretation must be defended through clear scientific reasoning or literature precedence in the paper.

Figure 2: Custom prompt given to Copilot

Rather than generating the entire paper in a single step, we employed an iterative reconstruction process. As a result, the agent was able to maintain context across the given iterative prompts. To prevent the generation process from being influenced by references or information from unrelated conversations, the agent's access to past chats was disabled. The process consisted of six stages described below.

The empirical-only manuscript was provided to the AI agent together with a prompt instructing the system to generate the *Related Work* section and *References* in Bibtex format, which were then inserted into the manuscript (Figure 3).

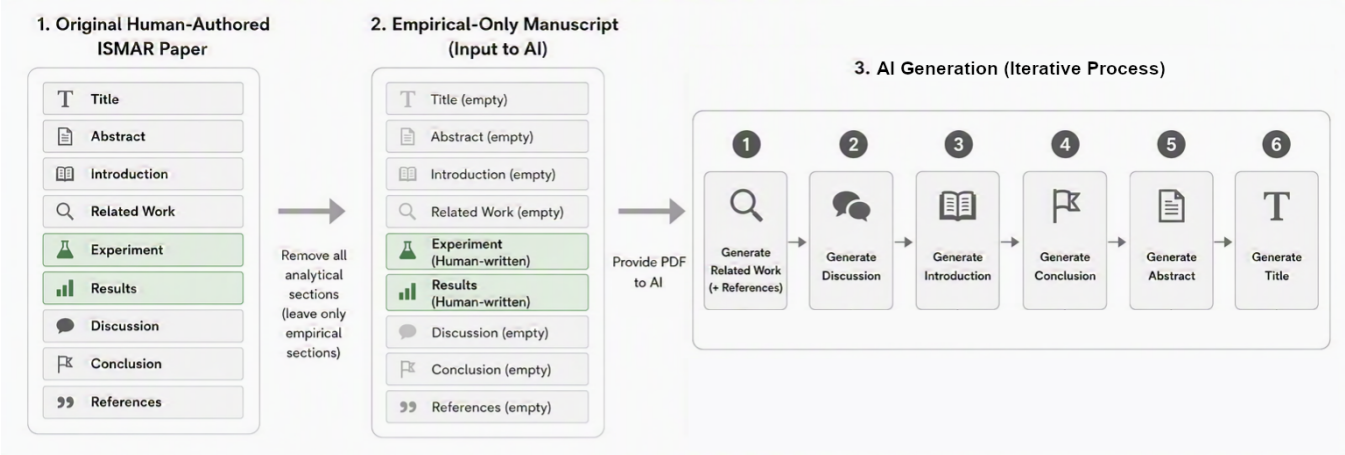

Figure 1: Illustration of the process we used. The empirical parts remain human-authored; everything else is generated by AI.

Prompt. Read through the attached document and understand the experiment description and results, both based on text and all the images in the draft. Generate a "Related Work" section in a format typically seen in accepted full papers at the ISMAR conference. Use only real, peer-reviewed sources, and verify every reference before including it by confirming the title, authors, publication venue, and year. Ensure that each citation directly supports the statement it accompanies, and do not include any reference that cannot be confidently validated. Provide accurate in-text citations and a complete reference list with no fabricated, duplicate, or unverifiable references. The generated section must be in context of the experiment and all references in Bibtex format from related VR/AR/HCI papers. Provide only the LaTeX code for the section content that can be directly copied and pasted into an existing LaTeX manuscript, using "\cite{}" commands for all in-text citations. Ensure there are no duplicate citations or BibTeX entries, and that every "\cite{}" command has a matching BibTeX entry.

Figure 3: Prompt given to the agent to generate the 'Related Work' section of the paper.

2.3.1 Stage II: Discussion Generation

Using the expanded context, Copilot was then instructed to generate a *Discussion* section interpreting the findings from the experiment and relating them to the identified literature.

Prompt. Using the manuscript, Related Work section, bibliography, and all previous inputs and outputs from this conversation as context, write the "Discussion" section of the paper. Do not request the manuscript again; continue from the information already provided and maintain consistency with the existing content and references. Provide only the LaTeX code for the Discussion section that can be directly copied and pasted into the manuscript. The Discussion should interpret the significant findings, explain their implications, and relate them to validated prior literature. All claims must be directly supported by the reported results or verified citations. Avoid unsupported speculation. Reuse existing references whenever possible. If new references are needed, add only new, peer-reviewed sources that are not already in the existing bibliography. Verify each new reference before including it and do not create duplicate citations or BibTeX entries. Use "\cite{}" for citations. After the LaTeX section, provide only the new BibTeX entries introduced in this section, excluding all previously generated references.

Figure 4: Prompt given to the agent to generate the 'Discussion' section of the paper.

2.3.2 Stage III: Introduction Generation

Using this information, we then used a prompt to ask Copilot to generate the *Introduction* section that motivated the study and positions it within the broader literature (Figure 5).

Prompt. Using the manuscript, Related Work section, Discussion section, bibliography, and all previous inputs and outputs from this conversation as context, write the "Introduction" section of the paper. Do not request the manuscript again; continue from the information already provided and maintain consistency with the existing content and references. Provide only the LaTeX code for the Introduction section that can be directly copied and pasted into the manuscript. The Introduction should establish the motivation, background, research gap, significance, and contributions of the work. Keep the introduction concise, approximately half a page or 4 to 6 paragraphs in a standard two-column format. Ensure all claims are supported by the manuscript findings or validated literature citations, avoiding unsupported or exaggerated statements. Reuse existing references whenever possible. If new references are needed, add only new, peer-reviewed sources that are not already in the existing bibliography. Verify each new reference before including it and do not create duplicate citations or BibTeX entries. Use "\cite{}" for citations. After the LaTeX section, provide only the new BibTeX entries introduced in this section, excluding all previously generated references.

Figure 5: Prompt given to the agent to generate the 'Introduction' section of the paper.

2.3.3 Stage IV: Conclusion Generation

We then asked the AI to generate a *Conclusion* section for the paper (Figure 6).

Prompt. Using the manuscript, Related Work section, Discussion section, Introduction section, bibliography, and all previous inputs and outputs from this conversation as context, write the "Conclusion" section of the paper. Do not request the manuscript again; continue from the information already provided and maintain consistency with the existing content and references. Provide only the LaTeX code for the Conclusion section that can be directly copied and pasted into the manuscript. The Conclusion should summarize the key findings, highlight the contributions of the work, discuss the broader implications, and briefly mention future directions if appropriate. Keep the conclusion concise, about one paragraph or approximately 200 words. Ensure all statements are supported by the reported results and avoid introducing new claims, unsupported interpretations, or new literature. Reuse existing references whenever possible. If new references are needed, add only new, peer-reviewed sources that are not already in the existing bibliography. Verify each new reference before including it and do not create duplicate citations or BibTeX entries. Use "\cite{}" for citations. After the LaTeX section, provide only the new BibTeX entries introduced in this section, excluding all previously generated references.

Figure 6: Prompt given to the agent to generate the 'Conclusion' section of the paper.

2.3.4 Stage V: Abstract Generation

Once all major manuscript components had been generated, Copilot was asked to produce an *Abstract* summarizing the paper (Figure 7).

Prompt. Using the manuscript and all previous outputs from this conversation as context, write the Abstract section. Maintain consistency with the existing content, terminology, research objectives, methodology, findings, and contributions. Provide only the LaTeX code for the Abstract that can be directly copied into the manuscript. Do not generate a complete LaTeX file. The Abstract should briefly introduce the research problem, describe the approach/methodology, summarize the key findings, and state the main contributions and implications of the work. Keep the abstract concise, about one small paragraph or approximately 200 words. Keep it concise, clear, and suitable for a peer-reviewed research paper. Do not include citations or references in the Abstract, and avoid introducing information not supported by the manuscript.

Figure 7: Prompt given to the agent to generate the 'Abstract' section of the paper.

2.3.5 Stage VI Title Generation

Once components that made up the paper had been generated, Copilot was asked to produce a *Title* summarizing the paper (Figure 8).

Prompt. Using the manuscript and all previous outputs from this conversation as context, generate a concise, publication-quality paper title. The title should accurately reflect the research problem, methodology, and key contributions while remaining clear and engaging. Aim for a title suitable for a top-tier academic conference or journal, avoiding overly generic, sensational, or unnecessarily long phrasing. Return only the title, with no explanations or additional text.

Figure 8: Prompt given to the agent to generate the title of the paper.

2.4 Preservation of AI Outputs

The AI outputs were inserted into the manuscript largely as generated so that the resulting document would accurately reflect the analytical framing produced by AI rather than a hybrid human-AI authorship process. The only modifications we made were twofold:

a. Minor adjustments to the LaTeX structural and style commands, which were not always consistently included in the AI-generated content, and

b. Consolidation of the sometimes duplicated references that were included by the AI in each step.

The objective was not to create the best possible AI-assisted paper, but rather to observe what scientific narrative emerges when analytical authorship and framing is delegated almost entirely to an AI system.

2.5 Comparative Analysis

The AI-framed paper was compared with the original human-authored version along three dimensions:

- *Structural Analysis:* We compared a few basic structural characteristics of the two manuscripts.
- *Automated Review Analysis:* Both papers were submitted to the same AI-based review system (*paperreview.ai* by Stanford ML Group).
- *Author Reflection:* Finally, we obtained qualitative reflections from the original authors of the human-authored paper about the AI-framed version.

3 Results and Discussion

In this section, we present and discuss the results of our analysis.

3.1 Structural Analysis

The two submissions were built from the same *Experiment* and *Results* sections, but differ substantially in structural organization. The AI-generated manuscript expands the *Related Work* into five subsections and ends its *Introduction* in a three-part contribution list. The human-written manuscript compresses *Related Work* into three subsections and structures its *Introduction* with three explicit research questions (RQ1–RQ3). The *Discussion* sections diverge most sharply: The AI version is partitioned into four hypothesis-labeled subsections (H1–H4) plus separate *Implications* and *Limitations* subsections, while the human-written version uses three conceptually-labeled subsections, with the *Direction of the Transition Matters* subsection further nested into two sub-subsections. The references also differ, with 22 entries in the AI version versus 47 in the human version.

3.2 Automated Review Analysis

We ran both versions of the paper through the AI-automated paper reviewing system *paperreview.ai* by Stanford ML Group, resulting in separate AI-generated reviews for the AI-framed paper[1] and the human-written paper[2].

Because the *Experiment* and *Results* sections were identical across both paper versions, differences in the resulting reviews primarily reflect differences in analytical framing rather than differences in the underlying empirical evidence. Both reviews identified largely the same methodological concerns. The raised concerns were directly tied to the experimental design and analysis and therefore appeared largely independent of how the study was framed. This suggests that the empirical evidence itself remained the primary source of methodological criticism regardless of whether the paper was written by humans or AI.

However, the reviews differed in their assessment of the paper's narrative structure. The review of the human-written paper praised the clarity of the motivation, research questions, and experimental design, and discussed the study as a well-executed empirical contribution. In contrast, the review of the AI-generated paper repeatedly highlighted the coherence of the proposed theoretical construct, the integration of the research literature, and the consistency of the overall narrative. The review appeared to regard the AI-generated paper as presenting a more unified conceptual contribution, despite relying on exactly the same empirical evidence.

Interestingly, the stronger theoretical framing of the AI-generated paper also resulted in different forms of criticism. While both reviews raised largely identical methodological concerns, the review of the AI-generated paper placed greater emphasis on whether the proposed modality-return cost was genuinely a modality effect or whether it could instead be explained by depth-switching, accommodation–vergence, or other perceptual factors. In other words, once the AI-generated paper promoted a more ambitious theoretical interpretation, the review began evaluating the work according to that interpretation and demanded stronger evidence in support of the proposed conceptual contribution.

Overall, the comparison suggests that the AI-generated framing influenced how the study was interpreted and evaluated. The original paper was reviewed primarily as a careful empirical XR study, whereas the AI-generated paper was reviewed as a more theory-driven contribution to modality-switching research. Most notably, the reviews appeared to fully adopt the AI-generated framing and terminology, treating the proposed "asymmetric modality-return cost" as a legitimate scientific contribution despite that concept not serving as the primary motivation for the original research.

[1] https://paperreview.ai/review?token=yyNn4TstjZa5DU0wGVKpb8dnnpkY1z9bmOUn6U38tU8

[2] https://paperreview.ai/review?token=lWBdYZxm8gHC7zmkVRjQhphvOsMwLcVa6ShvdkoU5Y8

This finding suggests that AI-generated analytical framing may influence not only how research is communicated, but also how its scientific contribution is ultimately perceived and assessed during the review process.

The overall tone of the AI reviews also differed in a subtle but meaningful way. Although neither review assigned an explicit score, both concluded with an overall assessment and recommendation. The review of the human-written paper emphasized methodological concerns and repeatedly framed these issues as obstacles for publication at a top-tier venue, ultimately recommending a "major revision or resubmission." In contrast, the review of the AI-generated paper adopted a noticeably more positive and optimistic tone. While it identified many of the same methodological limitations, these were presented more often as opportunities for strengthening an otherwise valuable contribution. The review repeatedly described the work as "timely," "well-executed," "meaningful," and having "strong potential for publication," concluding that the paper could become publication-ready through revisions. In other words, despite both reviews identifying largely the same weaknesses in the underlying experiment, the AI-generated framing appears to have shifted the overall evaluation from one focused on methodological shortcomings toward one focused on the potential significance and impact of the contribution. This observation is particularly noteworthy because the empirical evidence was identical in both manuscripts. The difference in reviewer tone therefore appears to originate primarily from differences in narrative framing, conceptual positioning, and the perceived strength of the paper's claimed contribution.

3.3 Author Reflection

After comparing the original human-written paper and the AI-generated paper, the authors concluded that the AI-generated paper represents a substantially different interpretation of the same empirical evidence. Although the *Experiment* and *Results* sections were identical in both manuscripts, the analytical framing diverged considerably.

The original paper reflected a traditional empirical research process in the field of XR in which research questions motivated the experiment and the resulting findings were interpreted within that context. Most notably, the authors discuss that the results from the study were contrary to their initial expectations. The AI-generated paper followed the opposite trajectory. It began with the observed findings and worked backwards to infer contributions to the field. As a result, some arguments appeared more like post-hoc interpretations than consequences of the original research agenda. The authors therefore viewed the AI-generated paper as less faithful to the analytical process that originally led to the experiment, though it is hard to argue that the resulting paper is less compelling in its framing of the empirical findings.

The human-written paper situated the work within prior AR research, emphasizing perceptual constraints, ambient lighting, visual fatigue, and display-switching behavior with AR displays. By comparison, the AI-generated paper organized the study around concepts from task-switching theory and repeatedly connected the observed findings to broader theoretical constructs. Most notably, the AI introduced the concept of an "asymmetric modality-return cost" and used it as the central organizing principle of the paper. This concept appeared consistently throughout the title, abstract, introduction, discussion, and conclusion, creating a focused narrative structure. In other words, the AI identified a single finding and elevated the context to the motivation for conducting this research and the central scientific contribution of the paper.

Perhaps the most surprising observation was that the AI-generated framing was arguably good enough (or close to it) for publication. The paper was coherent, well-structured, and aligned with the scientific language and conventions in this field. If evaluated independently, reviewers might not suspect that the analytical sections had been generated entirely by AI. Some reviewers might even prefer the AI-generated framing because of its narrative clarity and more theoretical focus. This observation reveals an important tension. The AI-generated paper was not a faithful representation of the original intellectual motivations behind the work, yet it arguably aligned with many of the qualities rewarded by an academic publication process: contributions, narrative, theoretical implications, and a compelling explanation of the findings.

Overall, the AI identified a plausible scientific contribution that the authors did not originally intend to make. Given identical empirical evidence, the human-written and AI-framed papers converged on different interpretations of what the study was about. This raises a broader question for the future of scientific publishing: If AI can generate a compelling narrative for an XR study, should scientific papers prioritize faithfulness to the original research agenda or optimization of the exposition for consumption by the research community in this field?

3.4 Limitations

This study has several limitations that should be considered when interpreting its findings. First, each section of the paper was generated only once, and no attempt was made to regenerate or iteratively refine the AI's responses. The AI agent was not given any hints about being right or wrong. Consequently, the evaluation reflects the model's initial output rather than its best possible performance. It is possible that alternative generations or iterative prompting could have produced responses that more closely resembled the human-authored manuscript [28] or produced an output based on the target of the experiments itself. Finally, this evaluation was conducted using a single empirical research paper and a single AI model. Different research domains, writing styles, prompting strategies, or language models may produce different outcomes. Future work should expand this analysis to multiple papers and AI systems to determine whether the trends observed in this study generalize across a broader range of scientific disciplines and writing tasks [9]. In addition, future studies could generate papers multiple times using different seed papers to examine how variations in the initial source material influence the resulting manuscripts, or incorporate qualitative feedback from the original authors to evaluate the perceived accuracy, quality, and usefulness of AI-generated manuscripts.

4 Conclusion

This paper explored how well the analytical portions of an empirical XR paper could be generated entirely by AI when provided only with the *Experiment* and *Results* sections. Using an accepted, but yet-to-be-published, ISMAR 2026 paper as a case study, we found that AI was able to generate a complete, coherent, and publication-ready manuscript whose framing, motivations, and implications differed substantially from those of the original human-written paper. Rather than reproducing the authors' original research narrative, the AI constructed a new narrative around the most salient empirical finding, producing a paper that may even be perceived as more persuasive, but less connected to the motivations that originally inspired the study. Our results suggest that AI-generated analytical framing may already appear feasible for empirical research to large portions of the XR research community and may soon become a routine part of scientific writing. The challenge for the research community is therefore no longer whether AI can write papers, but how to preserve researcher intent, intellectual honesty, and scientific accountability when compelling stories about empirical evidence can increasingly be generated by AI.

Acknowledgments

This material includes work supported in part by the National Science Foundation under Award Number 2235066 (Dr. Han-Wei Shen, IIS); the Office of Naval Research under Award Numbers N00014-25-1-2159 and N00014-25-1-2245 (Dr. Peter Squire, Code 34); and the AdventHealth Endowed Chair in Healthcare Simulation (Prof. Welch).

References

[1] A. Asai, J. He, R. Shao, W. Shi, A. Singh, J. C. Chang, K. Lo, L. Soldaini, S. Feldman, M. D'Arcy, et al. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, pp. 1–7, 2026. 1

[2] Z. D. Choudhary, L. Lisle, A. Alwin, G. Bruder, and G. Welch. Cognitive performance during display switching in optical see-through augmented reality. *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2026. 2

[3] G. Erol, A. Ergen, B. Gülşen Erol, Ş. Kaya Ergen, T. S. Bora, A. D. Çölgeçen, B. Araz, C. Şahin, G. Bostancı, İ. Kılıç, et al. Can we trust academic ai detective? accuracy and limitations of ai-output detectors. *Acta neurochirurgica*, 167(1):214, 2025. 1

[4] C. Gonzalez, K. Donahue, D. G. Goldstein, H. Heidari, M. S. Jalali, B. Schelble, A. Singh, and A. W. Woolley. Toward a science of human–ai teaming for decision making: A complementarity framework. *PNAS nexus*, 5(3):pgag030, 2026. 1

[5] M. Gottsacker, A. Alwin, H. Furuya, R. Lindeman, G. Bruder, and G. Welch. ReVoicer: Conversational Voice Annotation for Human-Centered, LLM-Assisted Peer Review. In *Alt'ISMAR: Adjunct Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 1–8, 2026. 1

[6] A. Gray. Chatgpt" contamination": estimating the prevalence of llms in the scholarly literature. *arXiv preprint arXiv:2403.16887*, 2024. 1

[7] S. P. Horbach, E. K. Schmidt, R. Fishberg, and M. P. Sørensen. Writing assistant, workhorse, or accelerator? how academics are using genai. 2024. 1

[8] M. Khalifa and M. Albadawy. Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer methods and programs in biomedicine update*, 5:100145, 2024. 1

[9] M. T. R. Laskar, S. Alqahtani, M. S. Bari, M. Rahman, M. A. M. Khan, et al. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13785–13816, 2024. 5

[10] G. F. Lendvai. Chatgpt in academic writing: A scientometric analysis of literature published between 2022 and 2023. *Journal of Empirical Research on Human Research Ethics*, 20(3), 2025. 1, 2

[11] X. Li and J. Ouyang. Automatic related work generation: A meta study. *arXiv preprint arXiv:2201.01880*, 2022. 1

[12] I. S. MacKenzie. *Human-Computer Interaction: An Empirical Research Perspective*. Elsevier, 2024. 1

[13] I. S. MacKenzie, J. C. Read, and M. Horton. Empirical research methods for human-computer interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–3. ACM, 2024. 1

[14] L. Nolte and S. Tomforde. A helping hand: a survey about ai-driven experimental design for accelerating scientific research. *Applied Sciences*, 15(9):5208, 2025. 1

[15] H. H. Rashidi, B. D. Fennell, S. Albahra, B. Hu, and T. Gorbett. The chatgpt conundrum: Human-generated scientific manuscripts misidentified as ai creations by ai text detection tool. *Journal of Pathology Informatics*, 14:100342, 2023. 1

[16] F. Şahinuç, S. Dutta, and I. Gurevych. Expert preference-based evaluation of automated related work generation. *arXiv preprint arXiv:2508.07955*, 2025. 1

[17] D. Scherbakov, N. Hubig, V. Jansari, A. Bakumenko, and L. A. Lenert. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *Journal of the American Medical Informatics Association*, 32(6):1071–1086, 2025. 1

[18] E. Shao, Y. Wang, Y. Qian, Z. Pan, H. Liu, and D. Wang. Sciscigpt: advancing human–ai collaboration in the science of science. *Nature Computational Science*, 6(3):301–315, 2026. 1

[19] P. Spitzer, J. Holstein, P. Hemmer, M. Vössing, N. Kühl, D. Martin, and G. Satzger. Human delegation behavior in human-ai collaboration: The effect of contextual information. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–28, 2025. 1

[20] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. V. Katwyk, A. Deac, A. Anandkumar, K. Bergen, C. P. Gomes, S. Ho, P. Kohli, J. Lasenby, J. Leskovec, T.-Y. Liu, A. Manrai, D. Marks, B. Ramsundar, L. Song, J. Sun, J. Tang, P. Veličković, M. Welling, L. Zhang, C. W. Coley, Y. Bengio, and M. Zitnik. Scientific discovery in the age of artificial intelligence. *Nature*, 620:47–60, 2023. 1

[21] Z. P. Wang, P. Bhandary, Y. Wang, and J. H. Moore. Using gpt-4 to write a scientific review article: a pilot evaluation study. *BioData Mining*, 17(1):16, 2024. 1

[22] S. Watson, E. Brezovec, and J. Romic. The role of generative ai in academic and scientific authorship: an autopoietic perspective. *AI & society*, 40(5):3225–3235, 2025. 1

[23] S. Wu, X. Ma, D. Luo, L. Li, X. Shi, X. Chang, X. Lin, R. Luo, C. Pei, C. Du, Z.-J. Zhao, and J. Gong. Automated literature research and review-generation method based on large language models. *National Science Review*, 12(6), 2025. 1

[24] Z. Xu, Y. Zhao, M. Patwardhan, L. Vig, and A. Cohan. Can llms identify critical limitations within scientific research? a systematic evaluation on ai research papers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20652–20706, 2025. 1

[25] Y. Yamada, R. T. Lange, C. Lu, S. Hu, C. Lu, J. Foerster, J. Clune, and D. Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025. 2

[26] Y. Zhang, S. A. Khan, A. Mahmud, H. Yang, A. Lavin, M. Levin, J. Frey, J. Dunnmon, J. Evans, A. Bundy, S. Dzeroski, J. Tegner, and H. Zenil. Exploring the role of large language models in the scientific method: From hypothesis to discovery. *npj Artificial Intelligence*, 2025. 1

[27] H. Zheng and H. Zhan. ChatGPT in Scientific Writing: A Cautionary Tale. *The American Journal of Medicine*, 136(8):725–726, 2023. 1

[28] J. Zhuo, S. Zhang, X. Fang, H. Duan, D. Lin, and K. Chen. Prosa: Assessing and understanding the prompt sensitivity of llms. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 1950–1976, 2024. 5