How useful is gen AI really? A comparative study of gen AI tools for academic writing in engineering education

Briana Lavine, Purdue University – West Lafayette (College of Engineering)
Dr. Adrian Nat Gentry, Purdue University – West Lafayette (College of Engineering)

Adrian Nat Gentry is a postdoctoral researcher in Engineering Education at Purdue University. They have a Ph.D. student in Engineering Education and a master's and bachelor's in Materials Engineering. Adrian's research interests include assessing student supports in cooperative education programs and the experiences and needs of nonbinary scientists.

Dr. Kerrie A Douglas, Purdue University – West Lafayette (College of Engineering)

Dr. Douglas is an Associate Professor in the Purdue School of Engineering Education. Her research is focused on improving methods of assessment in engineering learning environments and supporting engineering students.

Dr. Hannah Kim, Purdue University – West Lafayette (College of Engineering)

Hannah Kim is a postdoctoral researcher in Purdue University's School of Engineering Education. With a background in Learning Design and Technology, she develops scalable, evidence-based instructional systems that improve workforce readiness, learner engagement, and persistence in STEM and semiconductor workforce pathways.

How useful is gen AI really? A comparative study of gen AI tools for academic writing in engineering education

Introduction

This WIP research examines gen AI tools for their accuracy in literature review tasks for engineering education researchers. Following their rise in popularity in 2022, chat-based generative artificial intelligence tools (gen AI), such as ChatGPT, Copilot, and Gemini, have changed everyday engineering workflows across technical and educational sectors. Chat-based gen AI tools have increased ease of access to AI models; since 2023, industry organizations reported a 46% increase (33% to 79%) in their use of gen AI in at least one business function and a 33% increase (55% to 88%) in overall AI adoption [1].

Similar trends in popularity are evident across users in engineering education. Survey data from fall 2023 show that 70% of both STEM students and faculty have engaged with these tools multiple times a month [2]. For engineering students, gen AI usage is predominantly focused on academic and classroom-related tasks: gathering information (83%), summarizing content (73%), getting assignment assistance (63%), and generating content (60%) [3]. Although students recognize notable risks—such as inability to verify accuracy (65%), unreliable responses (62%), and unknown sources (50%)—the distinctive advantages gen AI tools continue to drive their adoption.

Among the various applications of gen-AI tools in teaching and learning, their use in conducting literature reviews has become especially prominent in higher education. These tools are reported to save time and reduce workload [4] while also supporting the generation of novel insights and facilitating more efficient interdisciplinary collaboration [5]. Prior work has examined specific aspects of AI-assisted research processes: Bolaños et al. [4] reviewed AI-enhanced tools available for systematic literature review and outlined guidelines for evaluation of the AI-generated outputs for performance, usability, and transparency. Hanafi et al. [5] proposed workflows that utilize multiple AI tools across different stages of the research process, emphasizing human oversight, transparency, and ethical considerations alongside efficiency gains. Similarly, Naghdy [6] described a structured approach for accelerating literature review and research ideation using gen-AI tools, demonstrating its potential for collaborative research through a case study.

While potentially valuable, the use of gen-AI for literature review presents a particularly high-stakes application, as literature synthesis forms the evidentiary foundation of scholarly inquiry and research design [7]. Fewer than 20% of gen AI-generated outputs are reviewed before use, according to industry professionals' self-reports [1]. Inaccurate summaries can propagate methodological misunderstandings, misrepresent empirical findings, and compromise the integrity of subsequent research. These risks are heightened for students and novice researchers who may lack the disciplinary expertise needed to detect subtle inaccuracies. While the accuracy of gen-AI outputs varies by subject area and question types [3], existing studies have focused primarily on usability, efficiency, workflow integration, and transparency, with limited attention to the accuracy or fidelity of AI-generated research syntheses. Given the rapid rise in reliance on

chat-based gen-AI systems for locating, summarizing, and synthesizing research, a systematic evaluation of their accuracy has become both timely and essential.

The purpose of this work-in-progress is to inform engineering education researchers and researchers-in-training about the extent to which popular chat-based tools can understand empirical engineering education literature, and where expert human interpretation remains necessary. We pilot our study by evaluating eight popular chat-based gen AI tools in terms of how accurate they are able to report out critical research information from two engineering education published papers, Holloway et al. [8] and Douglas et al. [9]. We were guided by the question: which tools are most accurate in synthesizing research findings for assessment development articles? This analysis aims not only to identify comparable strengths and limitations but also to highlight what academic researchers and students should attend to when reviewing gen AI-generated literature reviews. The findings can inform researchers and students in determining which tools are most trustworthy for synthesizing journal articles and in applying appropriate evaluative checks to ensure the accuracy and rigor of literature reviews.

Methods

Prompt iteration

We designed our prompt to both capture crucial components of a literature review or annotated bibliography and reflect how students and faculty may seek information. First, we identified five criteria, based on their relevance to literature reviews, as defined by Schryen et al. [10] as extracting and synthesizing research activity: research questions, purpose of the instrument, breakdown of sample demographics, summary of analyses performed, and summary of results. We then utilized five separate prompts to isolate the criterion. However, after discussion, we decided that isolated prompts were a less common prompting method for students because isolated prompts may exhaust interactions with free versions quicker and are more time consuming to write. Thus, we decided to combine the prompts to one clear, concise prompt: "Given the research paper, can you provide the research questions, purpose of the instrument being evaluated, breakdown of sample demographics, summary of analyses performed, and summarize the results from the text?" Additionally, we chose to not ask any follow-up questions to further imitate students' usage with the tools and took each response exactly as output by the tools.

Gen AI Tools Evaluated

We selected eight gen AI tools to evaluate their accuracy in extracting and synthesizing academic literature: ChatGPT-5 (OpenAI), Gemini 2.5 Flash (Google DeepMind), Claude Sonnet 4.5 (Anthropic), NotebookLM (Google Labs), Qwen3-Max (Alibaba Cloud), Le Chat (Mistral AI), Copilot (Microsoft), and Perplexity (Perplexity AI, Inc.). We focused our study on gen AI tools as opposed to other AI tools or LLMs because of their accessibility, ease of use, and ability to create original ideas. Tools were chosen either based on their popularity among students (e.g., Chat-GPT5), having free versions easily accessible to students (i.e., Copilot), or their gaining attention in the academic sphere being built to assist with tasks aligning with our needs (e.g., Qwen3-Max). Additionally, we focused our scope on tools that utilize chat-based interfaces rather than user interfaces due to the differences in their outputs. Specifically, chat-based interfaces respond to prompts in different ways than user interface-based tools making it

difficult for fair comparison. While this list of tools is far from exhaustive, this study aims to provide a basis for evaluating and selecting chat-based gen AI tools and open up avenues for future work exploring increasingly task-specified tools or non-chat-based tools.

Research papers

We utilized two assessment development manuscripts in our evaluation: *Research experiences instrument: Validation evidence for an instrument to assess the research experiences of engineering PhD students' professional practice opportunities* by Holloway et al. [8] and *The Critical-Thinking Engineering Information Literacy Test (CELT): A Validation Study for Fair Use Among Diverse Students* by Douglas et al. [9]. These papers followed similar formatting when presenting their methods and results which allowed comparison between the gen AI tools' responses after extracting data.

Evaluation of Accuracy

To ensure that our researcher responses were accurate and easily comparable, the research team took multiple steps to create accurate, human-derived synthesis. First, we ran the prompt through each tool for each paper and the outputs were compiled into a master Excel file. Then, the two papers were annotated with five different colors to represent the five criteria. When relevant information was mentioned in the paper, the sentence was highlighted accordingly to create a human derived synthesis for each criterion. For research question and sample demographics a predetermined answer was created. For purpose of instrument, summary of analyses, and summary of results, each response was evaluated on a case-by-case basis and corroborated with the text because of the complexity of answers and information available in the paper. For example, in both papers, the purpose of the instrument was stated in the introduction, methods, discussion, and conclusion, meaning the gen AI tools could extract information across each of these sections, so each section had to be checked to verify accuracy.

Each tool output for the five criteria was evaluated for accuracy. To best compare and evaluate the depth and accuracy of each answer, the team evaluated each output—with two team members evaluating—utilizing research expertise, annotations and memos on the paper, and references directly with the text. Outputs were marked as accurate, meaning responses were complete and fairly represented the text or directly utilized the text, and inaccurate, meaning responses were missing information, too vague, included misleading information, or inaccurate. Once each tool's responses were evaluated for the two papers and five criteria, the original determination was cross-referenced and discussed with other members of the team. This study has several limitations, including the small number of papers reviewed, the narrow scope of article type, and limited prompting approach.

Results

Gen AI tool accuracy for Douglas et al. [9]

Research question

All tools created accurate outputs for the research questions by quoting directly from the text. Some tools implemented single-word or phrase changes to the questions while remaining consistent to what the paper was conveying.

Purpose of Instrument
The tools accurately depicted the purpose of the instrument through bullet points or condensed summaries, and all information could be traced back to distinct text in the paper.

Demographics
None of the tools were able to accurately and fully output this criterion. The tools presented inaccurate information such as computing wrong information (e.g., Perplexity stated a self-derived, assumed number instead of the given percentage), expressing misleading information, or omitting certain parts of the demographics that are crucial to the study. Five out of eight tools gave misleading information because they interpreted the pre-data cleaning sample as the sample used for analysis. Six out of eight tools omitted the demographic breakdown used for the DIF analysis.

Analyses
Half of the tools gave complete and correct synthesis of the analyses. The other half of tools made common mistakes by omitting crucial analyses completed (e.g., reliability testing) and misinterpreting the sample groups (e.g., demographic for DIF) who were analyzed in different sections of the study.

Results
The only tool to accurately output the results with all the necessary details was ChatGPT-5. With Copilot, there were no explicit errors, but the description was vague and lacking detail to the point that the results were not properly portrayed. For the other six tools, they misinterpreted the data analysis leading to incorrect synthesis.

Table 2. Accuracy of Tools for Douglas et al. (2018)

Prompt / **Tool**	RQ	Purpose of Instrument	Sample Demographics	Analyses	Results	Total accurate (out of 5)
ChatGPT-5	✓	✓	X	X	✓	3
Gemini 2.5 Flash	✓	✓	X	✓	X	3
Claude Sonnet 4.5	✓	✓	X	✓	X	3
NotebookLM	✓	✓	X	✓	X	3
Qwen3-Max	✓	✓	X	X	X	2
Le Chat	✓	✓	X	X	X	2
Copilot	✓	✓	X	✓	X	3
Perplexity	✓	✓	X	X	X	2

Gen AI tool accuracy for Holloway et al. [8]
Research question
The gen AI tools that performed well either output the text verbatim or made slight changes without altering the meaning of the question. One tool, Copilot, inaccurately stated the research

question because it omitted and added its own terms ultimately changing the purpose of the question.

Purpose of Instrument

Almost all the tools provided an accurate, condensed explanation of the purpose of the instrument by extracting words directly from the text. However, ChatGPT-5 gave a description of the purpose of the study, but it failed to mention the purpose of the instrument.

Demographics

All the tools except for one, Qwen3-Max, failed to provide accurate, complete information regarding the demographics in the paper. The tools made similar mistakes in failing to include crucial information, such as engineering disciplines or the sample breakdowns for all phases of the study.

Analyses

Most of the tools (6/8) provided accurate detailing of the analyses allowing for full understanding of how and why the analyses were used. Two tools, Perplexity and ChatGPT-5, missed important analyses (e.g., EFA and parallel analysis) leading to incomplete synthesis of the analyses.

Results

Five out of eight tools were able to accurately and completely output the results. Each of the tools' outputs could be tracked directly to sentences in the article. The tools that failed this criterion, Copilot, Le Chat, and ChatGPT-5, had common issues with accuracy because they provided too little detail or overgeneralized the results to the point of being misleading. Another issue seen with ChatGPT-5 was hallucinating an interpretation that was not accurate or representative of the text.

Table 3. Accuracy of Tools for Holloway et al. (2022)

Prompt / **Tool**	RQ	Purpose of Instrument	Sample Demographics	Analyses	Results	Total accurate (out of 5)
ChatGPT-5	✓	X	X	X	X	1
Gemini 2.5 Flash	✓	✓	X	✓	✓	4
Claude Sonnet 4.5	✓	✓	X	✓	✓	4
NotebookLM	✓	✓	X	✓	✓	4
Qwen3-Max	✓	✓	✓	✓	✓	5
Le Chat	✓	✓	X	✓	X	3
Copilot	X	✓	X	✓	X	2
Perplexity	✓	✓	X	X	✓	3

Discussion

Generally, we found a lack of consistent accuracy across tools, but consistency across criteria—that is, tools performed well on largely extractive elements but were unreliable on synthesis-heavy components. For example, in Douglas et al. [9], there was little to no consistency in the tools' performance within the paper with each tool answering approximately half accurately, but there was consistency for four of the criteria (except for summary of analyses). Tools frequently outputted accurate responses for criteria requiring "cut and paste" answers with little synthesis (e.g., research questions and purpose of instrument). Errors clustered when information was dispersed across sections or required careful distinctions, producing plausible but misleading summaries. For example, within sample demographics, all tools failed to give accurate outputs because of the amount of information given throughout the paper on demographics and not explicitly stating the breakdown in one place.

When compared across papers, we found interesting trends in accuracy: the tools overall performed worse with Douglas et al. [9] compared to Holloway et al. [8]. One possible explanation for this difference is variation in writing style and overall manuscript quality. Holloway et al. [8] explicitly stated the criteria frequently and clearly throughout the paper whereas the Douglas et al. [9] had less repetition and various levels of detail and clarity, allowing for misinterpretation. Additionally, Holloway et al. [8] was published in a flagship journal, which may indicate that the paper had higher quality of writing or went through more rounds of revision making it easier for the tools to search and synthesize accurate responses. The stronger performance on Holloway et al. [8] than Douglas et al. [9] further suggests that manuscript clarity and repetition may materially shape tool accuracy, with clearer signposting enabling more faithful extraction and synthesis.

These findings add an accuracy-focused complement to prior work that has largely emphasized efficiency and workflow guidance for AI-assisted literature reviewing, e.g., [4], [5], [6]. Consistent with those cautions, our results show that AI-generated summaries are not necessarily correct, reinforcing the need for explicit human oversight, especially given how often AI outputs neglected for further evaluation [1].

Conclusion

While literature reviews involve both technical and higher-order creative tasks [11] our findings suggest that the eight gen AI tools we examined produced adequate outputs for primarily technical, extractive tasks but were far less reliable for nuanced tasks that require higher-order cognition (e.g., findings and analysis). With the current capabilities and training of the version of gen AI tools utilized, the resulted outputs were limited in nuance and depth needed to critically review and evaluate empirical engineering education literature for research purposes. Moving forward, we encourage researchers to systematically evaluate gen AI tools using relevant criteria, such as the criteria utilized for this study, before utilizing outputs for their own research. However, until gen AI tools are more developed for the nuance needed to synthesize literature for research tasks, it is essential that researchers read and critically evaluate literature to advance their research. Simply, by utilizing these tools without greater evaluation, engineering education researchers may not be building the expertise needed to understand the potential scholarship gaps

that can make genuine contributions to the engineering education field. We encourage researchers to consider when use is appropriate for their own expertise development and when use is appropriate to evaluate the tool and verify important findings. Future studies should evaluate a larger and more diverse set of manuscripts, test multiple prompt designs, and examine how tool performance varies by disciplinary writing conventions and statistical complexity.

References

[1] A. Singla, A. Sukharevsky, L. Yee, M. Chui, B. Hall, and T. Balakrishnan, "The state of AI in 2025: Agents, innovation, and transformation.," Quantum Black, 2025. [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[2] J. Kim *et al.*, "Examining faculty and student perceptions of generative AI in university courses," *Innov. High. Educ.*, pp. 1–33, 2025.

[3] H. Akolekar *et al.*, "The role of generative AI tools in shaping mechanical engineering education from an undergraduate perspective," *Sci. Rep.*, vol. 15, no. 1, p. 9214, 2025.

[4] F. Bolaños, A. Salatino, F. Osborne, and E. Motta, "Artificial intelligence for literature reviews: opportunities and challenges," *Artif. Intell. Rev.*, vol. 57, no. 10, p. 259, Aug. 2024, doi: 10.1007/s10462-024-10902-3.

[5] A. M. Hanafi, M. M. Al-mansi, O. A. Al-Sharif, and others, "Generative AI in Academia: A Comprehensive Review of Applications and Implications for the Research Process.," *Int. J. Eng. Appl. Sci.-Oct. 6 Univ.*, vol. 2, no. 1, pp. 91–110, 2025.

[6] F. Naghdy, "Collaboration with GenAI in engineering research design," *Data Knowl. Eng.*, vol. 159, p. 102445, 2025, doi: https://doi.org/10.1016/j.datak.2025.102445.

[7] L. A. Maggio, J. L. Sewell, and A. R. Artino Jr, "The literature review: A foundation for high-quality medical education research," *J. Grad. Med. Educ.*, vol. 8, no. 3, pp. 297–303, 2016.

[8] E. A. Holloway, K. A. Douglas, D. F. Radcliffe, and W. C. Oakes, "Research experiences instrument: Validation evidence for an instrument to assess the research experiences of engineering PhD students' professional practice opportunities," *J. Eng. Educ.*, vol. 111, no. 2, pp. 420–445, 2022.

[9] K. A. Douglas, T. M. Fernandez, S. Purzer, M. Fosmire, and A. Van Epps, "The critical-thinking engineering information literacy test (CELT): A validation study for fair use among diverse students," *Int. J. Eng. Educ.*, vol. 34, no. 4, pp. 1347–1362, 2018.

[10] G. Schryen, M. Marrone, and J. Yang, "Exploring the scope of generative AI in literature review development," *Electron. Mark.*, vol. 35, no. 1, p. 13, Feb. 2025, doi: 10.1007/s12525-025-00754-2.

[11] G. Tsafnat, P. Glasziou, M. K. Choong, A. Dunn, F. Galgani, and E. Coiera, "Systematic review automation technologies," *Syst. Rev.*, vol. 3, no. 1, p. 74, 2014.