VIETNAM JOURNAL OF EDUCATION, 2025
Volume 9, Issue 2, 253-264
https://doi.org/10.52296/vje.2025.481

ORIGINAL ARTICLE

Evaluating the Feasibility of Using Generative AI for Educational Research within the Context of Vietnam's 2018 General Education Curriculum

Ngoan Quynh Thi Nguyen[1,+], Ngoc My Tran[1], Hang Thu Vu[2], Lan Thi Tran[1], Ha Viet Thi Nguyen[1], Thuy Thanh Bui[1], Trinh Thanh Nguyen[1]

[1]The Vietnam National Institute of Educational Sciences, Vietnam;
[2]Oxford University, United Kingdom
[+]*Corresponding author* ● *Email: ngoannt@gesd.edu.vn*

Article history
Received: 07 October, 2024
Accepted: 24 June, 2025
Published: 30 June, 2025

Keywords
Generative AI, educational research, Vietnam's general education, Vietnam's 2018 General Education Curriculum

ABSTRACT
This study examines the feasibility of using generative AI tools in educational research within the context of Vietnam's 2018 General Education Curriculum. The research evaluates three AI-powered tools – ChatGPT, Gemini, and Copilot - amid growing interest in AI's integration into academic fields, particularly in education. The focus is on their strengths across specific areas of educational research: curriculum development, implementation requirements, and evaluation and assessment. The tools' performance is assessed based on five criteria: accuracy, comprehensiveness, logical clarity, relevance, and currency of information. ChatGPT performs effectively in global citizenship education (curriculum development) while Gemini excels in history assessment standards (evaluation and assessment). Copilot shows promise but struggles with accuracy in certain domains. Despite variations in performance, all tools demonstrate potential in improving research processes, especially in tasks where absolute precision is not critical. However, accuracy remains a significant challenge across all platforms. The findings suggest that AI tools can greatly enhance academic work when used with proper verification and structured commands, underscoring their practical applications and future potential in transforming research methodologies.

1. INTRODUCTION

In recent years, artificial intelligence (AI) has evolved exponentially and reshaped the way we receive and exchange information. One of the most prominent areas is Generative Artificial Intelligence (GAI) - a branch of AI designed to create new content, including text, images, videos, and audio, which has created a global sensation and having a far-reaching impact on various fields, including education (Lim et al., 2023).

The most prominent generative AI tools currently applied in educational research can be exemplified by ChatGPT (OpenAI), Gemini (Google), and Copilot (Microsoft), each offering distinct capabilities for supporting academic tasks. These tools leverage large language models to assist with idea generation, content creation, literature synthesis, and even data analysis (Nguyen-Trung et al., 2023; Songkram et al., 2024). ChatGPT, launched by OpenAI, is widely recognized for its ability to summarize complex concepts, generate structured research outputs, and act as a virtual assistant throughout the research process (Brown et al., 2020; Xames & Shefa, 2023; Yenduri et al., 2023). Gemini, developed by Google as a successor to Bard, integrates multimodal processing and is designed to handle various

types of input such as text, code, and images, making it particularly useful for educational and research applications (Hochmair et al., 2024; Nikolic et al., 2024; Rane et al., 2024). Meanwhile, Copilot, introduced by Microsoft in 2023, integrates with Microsoft 365 and offers unique features such as academic citation support and source-linked footnotes, enhancing its practicality in research settings (Camillepack, 2024; KelliDavis, 2024; Rossettini et al., 2024). Together, these tools represent a major advancement in integrating AI into educational research and practice.

As their influence grows, an increasing number of studies have sought to examine how generative AI can be integrated into research workflows, from literature review and idea generation to data analysis and manuscript preparation. A systematic mapping review of 407 publications on generative AI in education and research identifies eight discursive themes, predominantly focused on applications and impacts, ethical implications, user perspectives, institutional adoption, and performance evaluation (Yusuf et al., 2024). However, most existing research has been conducted in Western contexts, with limited exploration of the applicability and challenges of AI integration in non-Western or developing educational systems.

In the context of Vietnam's general education system, access to effective research tools is particularly critical. The 2018 General Education Curriculum reforms prioritize evidence-informed policy and practice, yet limited access to academic databases and uneven technological proficiency among educators hinder research productivity. In this context, generative AI tools could serve as a valuable support mechanism - offering assistance in generating ideas, summarising literature, and drafting initial analyses - thus narrowing the gap between research demands and available resources.

To date, however, little research has systematically evaluated the feasibility of using generative AI tools in educational research within Vietnam, particularly in relation to the 2018 curriculum reform. This gap highlights the need for a critical, context-specific investigation into how tools like ChatGPT, Gemini, and Copilot perform when applied to core areas of Vietnamese educational research.

Therefore, this study aims to explore the feasibility of integrating generative AI tools into the educational research process in Vietnam. By selecting these three areas as the basis for comparing and evaluating the feasibility of Generative AI tools such as ChatGPT, Copilot, and Gemini within the new educational context in Vietnam. The research aims not only to provide a comparative analysis of the feasibility of these three tools but also to offer specific recommendations for researchers, teachers, and policymakers on integrating AI technology into educational research. The findings from this study will contribute to enhancing the overall quality of educational research, helping the Vietnamese education system meet the demands of innovation and international integration.

2. LITERATURE REVIEW

2.1. Applications of Generative AI in Academic Research

Supporting Idea Generation and Research Planning: Generative AI, like ChatGPT, is an effective tool for generating initial ideas and outlining research topics (Rahman et al., 2023). It can be a useful tool for developing initial research ideas and creating research proposals. Academic researchers can leverage ChatGPT to get preliminary ideas on how to design their methodology sections. Songkram et al. (2024) also emphasize that ChatGPT has shown the potential to revolutionize academic research through its ability to generate creative ideas.

Supporting the Research Process: Many researchers and education experts indicate that ChatGPT is capable of producing a complete research paper, depending on the level of detail required by the user. According to Songkram et al. (2024), ChatGPT has been employed at different stages of the research process, from literature review to data analysis. For example, ChatGPT can assist in adjusting search queries from one database to another, saving time and improving efficiency in data synthesis (Nguyen Trung et al., 2023). For qualitative data, GPT-4 can work on multiple codes and clusters to generate themes although this process requires close supervision by researchers to avoid errors such as code duplication or misclassification (Nguyen Trung, 2024). Notably, ChatGPT can generate survey questions or questionnaires for research purposes, including various types of questions such as multiple-choice, open-ended, dichotomous, and scale questions, which offers flexibility in research design (Xames & Shefa, 2023).

Supporting Report Writing and Publication: AI will soon be capable of writing complete articles, conducting peer reviews, and assisting editorial boards in accepting or rejecting manuscripts (Van Dis et al., 2023). ChatGPT can not only simplify the report-writing process but also enrich the content, enabling researchers to share their insights more effectively. It thus helps enhance their ability to contribute to the broader academic community (Nguyen Trung

et al., 2023). Another significant benefit of ChatGPT is its ability to remove language barriers, which is particularly useful for researchers who do not use English as their primary language. ChatGPT can help them draft high-quality academic texts, thus enhancing the global reach and accessibility of their research (Hosseini et al., 2023; Xames & Shefa, 2023). Additionally, after preparing the manuscript, ChatGPT can act as a tool to suggest suitable journals for submission, providing recommendations based on the title and abstract of the draft, which saves time for researchers (Xames & Shefa, 2023).

2.2. Limitations of Generative AI in Academic Research

Lack of Citations and Accountability: ChatGPT is a potential research support tool but cannot be recognized as an author or co-author due to a lack of accountability (Peres et al., 2023). "*If ChatGPT deserves authorship, then Microsoft Word also deserves it for providing us with a platform to organize and write documents more efficiently... Excel, R, or Python deserve co-authorship for calculating statistics or analyzing data for a quantitative scientific publication*" (Karim, 2023, p. 5). In addition, ChatGPT tends to reproduce text without appropriate citations or acknowledgements, which can be challenging for researchers and raise concerns about plagiarism (Xames & Shefa, 2023).

Lack of Reliability: According to Van Dis et al. (2023), a test on ChatGPT's accuracy in synthesizing key findings in research fields found that ChatGPT often generates incorrect and misleading text. When asked to summarise a specific review article, "*ChatGPT fabricated a response containing factual errors, misrepresentations, and false data*" (Van Dis et al., 2023, p. 224). Similarly, when it is asked to generate citations, ChatGPT may occasionally introduce errors by providing inaccurate or entirely non-existent references (Xames & Shefa, 2023). ChatGPT may also experience "hallucinations," where it generates unreasonable or illogical responses, particularly in prolonged or complex conversations (Lakshmanan, 2022; Morgan, 2023). This diminishes the reliability of AI tools in supporting research. Furthermore, although ChatGPT has been trained on a large dataset of text, it does not have real-time access to external databases. This means that the information ChatGPT provides may not be up-to-date or relevant to the present context (Songkram et al., 2024). Since ChatGPT is trained on existing data samples, it may reinforce biases present in the data, leading to distorted research outcomes (Songkram et al., 2024). There is also a concern that the development of ChatGPT and similar tools may lead to an increase in pseudo-scientific content in academic literature if not strictly controlled (Xames & Shefa, 2023).

Lack of Consistency in Reasoning: The lack of transparency in AI's reasoning process makes it difficult for researchers to fully understand how results are generated (Songkram et al., 2024). In the case of ChatGPT, for example, the same input prompt may yield different outputs on different occasions (Megahed et al., 2023). The steps and prompts during the use of ChatGPT often need to be adjusted multiple times to achieve the desired result, reducing the consistency of the outputs that this tool can provide (Nguyen-Trung, 2024). The generation of inconsistent results is one of the major issues with these GAI tools (Nguyen-Trung, 2024).

Limitations in Information Processing Capabilities: Natural Language Processing (NLP) systems like ChatGPT mainly rely on statistical relationships between words in text without understanding the relationship between language and external reality. This means these systems may struggle to draw accurate conclusions or perform common-sense reasoning, potentially leading to the generation of incorrect or illogical arguments (Hosseini et al., 2023). Additionally, GAI models face many limitations when performing detailed tasks on large datasets, especially in qualitative data analysis (QDA) (Nguyen-Trung, 2024). They also encounter difficulties when dealing with abstract topics (Morgan, 2023). Another point to consider is that ChatGPT is not trained on specialized data directly related to our research fields or the fundamentals of educational research (Nguyen-Trung, 2024). While Generative AI tools show some understanding of research concepts, they may misinterpret the literature or produce misleading descriptions and summaries of these key concepts. Current AI tools cannot fully grasp complex concepts in research and require close human supervision to ensure accuracy (Nguyen-Trung, 2024).

2.3. The Application of AI in Vietnam's General Education Context

Vietnamese education is currently undergoing a comprehensive reform process, with the 2018 General Education Curriculum marking a significant shift toward competency-based, student-centered, and integrated learning (MOET, 2018). As part of this transition, educational research plays a crucial role in informing curriculum development, instructional practices, and assessment models aligned with the new vision.

In this context, the integration of Artificial Intelligence (AI) into educational research holds transformative potential. AI tools, particularly generative AI such as ChatGPT, Gemini, and Copilot, offer significant advantages,

including faster access to information, support for literature synthesis, assistance with drafting, and the ability to generate content tailored to research prompts (Nguyen Trung et al., 2023; Rahman et al., 2023; Songkram et al., 2024). These features can be especially valuable in resource-constrained environments, where researchers often lack access to robust databases, peer collaboration, or time for extensive data analysis.

Despite AI's promise to transform educational research, there is currently a lack of systematic studies evaluating its feasibility in the Vietnamese education context, particularly within the general education sector. Little comprehensive review or official guidance currently exists on how to effectively use generative AI tools in alignment with the demands and characteristics of the 2018 General Education Curriculum.

This study thus aims to examine the feasibility of using new AI technology in the research process in Vietnam. The three applications juxtaposed include ChatGPT 3.5, Copilot (Microsoft), and Gemini (Google). These three applications were chosen for this study because they are among the most prominent and widely recognized generative AI tools currently available to the public. This research investigates the feasibility of integrating generative AI technologies into the research process within the Vietnamese educational context, particularly following the launch of the 2018 General Education Curriculum.

3. MATERIALS AND METHODS

This study aims to evaluate the feasibility of generative AI tools for conducting educational research in the context of Vietnam's 2018 General Education Curriculum. The selected AI tools - ChatGPT (OpenAI), Gemini (Google), and Copilot (Microsoft) - would be used to generate content for three research areas relevant to the curriculum reform: (1) classroom facilities that meet the curriculum requirements, (2) curriculum development, and (3) evaluation and assessment. The methodology has been designed to ensure a rigorous, comparative assessment of AI-generated content using a combination of quantitative and qualitative approaches.

Research Design and Justification: The study employs a comparative case study design, wherein each of the three selected educational research areas serves as a case. Comparative case studies are particularly useful in this context as they enable in-depth exploration of multiple cases (Yin, 2018). By focusing on three different educational topics - each aligned with key elements of the 2018 General Education Curriculum - the study can provide a nuanced analysis of how generative AI tools perform across varied types of educational research. Each AI tool will be applied to the same research questions, enabling a structured comparison across outputs.

Selection of Educational Research Areas: The areas selected for this study - classroom facilities, curriculum development, and evaluation and assessment - are central to the Vietnam 2018 General Education Curriculum. Each area represents a distinct aspect of educational reform and is critical to understanding how the framework impacts the overall education system. The first area, classroom facilities, looks at the physical and logistical requirements necessary to implement the curriculum. Curriculum development, the second area, is a core focus of the reform, emphasizing student-centered learning and competency-based education (MOET, 2018). Finally, evaluation and assessment represent the means by which student progress is measured, a key concern in any education system. The decision to focus on these three areas aligns with prior research indicating that facilities, curriculum, and assessment are among the most critical elements influencing educational outcomes in developing contexts (UNESCO, 2016).

Data Collection: To generate the data, each generative AI tool - ChatGPT, Gemini, and Copilot - is given a prompt based on the research question for each educational area. The prompts would be crafted to ensure consistency, asking the AI tools to generate content that would theoretically serve as a foundation for an academic paper in each of the three research areas. Prompts will include contextual information about the 2018 General Education Curriculum to guide the AI's responses and ensure relevance.

The use of multiple AI tools reflects the growing interest in understanding the unique capabilities and limitations of different AI systems (Bender et al., 2021). Each of the selected tools has been chosen for its prominence in the field of natural language processing and content generation. ChatGPT, for instance, has been widely recognized for its conversational abilities and knowledge of a broad range of topics (OpenAI, 2023). Gemini, developed by Google, is built on deep learning models known for their focus on search and information retrieval (Google, 2023). Copilot, meanwhile, is known for its integration with productivity software and ability to assist in content creation (Microsoft, 2023). By including these tools, the study captures a range of generative AI capabilities and provides a well-rounded analysis.

Evaluation Instrument and Justification: The feasibility assessment framework in this study is developed based on evaluation frameworks from previous research, focusing on assessing the effectiveness of AI models such as ChatGPT, Copilot, and Gemini in education and medicine (Kung et al., 2023; Yıldız, 2023; Alasker et al., 2024; Gibson et al., 2024; Shang et al., 2024). Based on these frameworks, the researchers devised a new evaluation framework to measure the feasibility of AI models in the context of Vietnamese general education, particularly with the 2018 General Curriculum Framework. The evaluation criteria include:

(1) *Accuracy*: This criterion assesses the accuracy of the information provided by AI based on scientific validity (Kung et al., 2023; Yıldız, 2023; Alasker et al., 2024; Gibson et al., 2024). In the current study, when AI provides information related to content that has been researched and verified in the researcher's original article, the original article is considered the standard for comparison and evaluation. In cases where the AI presents information not available in the original article, the accuracy will be evaluated based on the reliability and authenticity of current scientific information;

(2) *Comprehensiveness*: This criterion evaluates the scope of the response, determining whether AI can provide all necessary aspects of the question (Yıldız, 2023; Alasker et al., 2024; Gibson et al., 2024; Shang et al., 2024);

(3) *Relevance*: Relevance is assessed based on the degree to which AI's responses address the core of the posed question (Kung et al., 2023; Shang et al., 2024);

(4) *Logicality and Clarity*: Based on scales from previous studies, this criterion measures the coherence and clarity of the information. Evaluating coherence and clear presentation ensures that the information is conveyed in an understandable and non-confusing manner (Gibson et al., 2024; Shang et al., 2024);

(5) *Up-to-date*: This criterion assesses AI's ability to provide updated information that reflects the latest developments in the relevant field (Shang et al., 2024). This is a critical factor in the context of education in Vietnam, with the 2018 General Education Curriculum frequently updated with new guidelines and information.

Criteria	Description	1	2	3	4	5
Accuracy	This criterion assesses the accuracy of the information provided by AI based on scientific validity. In the current study, when AI provides information related to content that has been researched and verified in the researcher's original article, the original article is considered the standard for comparison and evaluation.	Misinformation with numerous serious scientific errors	Some accurate information but many errors or uncertainty	Accurate information but may lack depth or be misunderstood in minor details	Most information is accurate and in-depth, with only minor errors that do not affect the overall outcome	Completely accurate information, without any errors or contradictions, and clearly presented
Comprehe-nsiveness	This criterion evaluates the scope of the response, determining whether AI can provide all necessary aspects of the question.	The answer completely lacks important aspects of the question, is incomplete or contradictory	The answer addresses some aspects of the question but is missing key elements	The answer provides basic information but lacks some extended or deeper details	The answer is complete, covering most necessary aspects with reasonable elaboration	The answer is comprehensive and thoroughly covers all aspects of the question with depth and no omissions
Relevance	This criterion is assessed based on	The information is	The information has	The information is	The information is	The information is

	the degree to which AI's responses address the core of the posed question.	unrelated or barely related to the question, failing to address the main point	some relevant elements but is largely unrelated or off-topic	relevant to the question but lacks detail or is incomplete in some areas	mostly relevant, covering most necessary aspects, with only minor details missing that do not significantly affect overall relevance	completely relevant to the question, fully and thoroughly covering all related aspects in detail
Logicality and Clarity	This criterion measures the coherence and clarity of the information. Evaluating coherence and clear presentation ensures that the information is conveyed in an understandable and non-confusing manner.	The information is illogical, disorganized, and causes significant confusion	The information is disorganized and may cause confusion or lack clarity	The information is logically structured and clear at a basic level but may lack detail or coherence in some parts	The information is clearly presented, logical, and easy to understand, with only minor errors that do not affect overall coherence	The information is completely clear, coherent, and easy to understand, with no errors or unclear points
Up-to-date	This criterion assesses AI's ability to provide updated information that reflects the latest developments in the relevant field (Shang et al., 2024). This is a critical factor in the context of education in Vietnam, with the 2018 General Curriculum Framework frequently updated with new guidelines and information.	The information is completely outdated, failing to reflect the latest events, knowledge, or developments in the field.	The information is mostly outdated, with only a few elements reflecting recent updates or changes.	The information is reasonably accurate but may miss significant updates or reflect only part of recent changes.	The information is mostly up-to-date, reflecting most of the recent changes and developments in the field, with only a few minor points lagging behind current trends.	The information is fully up-to-date, accurately reflecting all recent developments, events, and the latest knowledge in the field.

Data Analysis: Following content generation, the outputs will be evaluated by the original authors of the peer-reviewed papers on which the AI-generated content is based. These experts are best positioned to evaluate the alignment of AI-generated content with both the academic rigor of the original papers and the curriculum's objectives. The evaluators include senior researchers from the Vietnam National Institute of Educational Sciences and officials from the Ministry of Education and Training (MOET). All of the evaluators have extensive experience in educational research, particularly in curriculum development, assessment standards, and instructional design. Their expertise ensures that the evaluation process is grounded in domain-specific knowledge and reflects current educational policies and practices in Vietnam.

4. RESULTS AND DISCUSSIONS

4.1. Results

The following tables summarise the performance outcomes for three prominent AI-powered tools: ChatGPT, Gemini, and Copilot.

Table 1. Evaluation results of ChatGPT (OpenAI) across three research fields based on established criteria

Criteria	(1)	(2)	(3)	Average
Accuracy	3	2	3	**2.6**
Comprehensiveness	4	2	3	**3**
Relevance	4	3	3	**3.3**
Logicality and Clarity	4	3	3	**3.3**
Up-to-date	4	2	4	**3.3**
Average	**3.8**	**2.4**	**3.2**	

**Notes: (1) Global Citizenship Education in Ethics Education at Primary Education Level Within The 2018 General Education Curriculum; (2) Proposal for a Model of Physics Classrooms in High Schools to Meet the Requirements of the 2018 General Education Curriculum; (3) Research on Developing Competency Assessment Standards for High School Students in the History subject according to the 2018 General Education Curriculum*

Table 2. Evaluation results of Gemini (Google) across three research fields based on established criteria

Criteria	(1)	(2)	(3)	Average
Accuracy	2	2	3	**2.3**
Comprehensiveness	2	3	3	**2.6**
Relevance	2	3	3	**2.6**
Logicality and Clarity	3	3	3	**3**
Up-to-date	3	2	3	**2.6**
Average	**2.4**	**2.6**	**3**	

Table 3. Evaluation results of Copilot (Microsoft) across three research fields based on established criteria

Criteria	(1)	(2)	(3)	Average
Accuracy	2	1	3	**2**
Comprehensiveness	2	2	3	**2.3**
Relevance	2	2	4	**2.6**
Logicality and Clarity	2	2	3	**2.3**
Up-to-date	3	3	4	**3.3**
Average	**2.2**	**2**	**3.4**	

Table 4. Average results of three AI tools based on established criteria

Criteria	ChatGPT (Open AI)	Gemini (Google)	Copilot (Microsoft)
Accuracy	2.6	2.3	2
Comprehensiveness	3	2.6	2.3
Relevance	3.3	2.6	2.6
Logicality and Clarity	3.3	3	2.3

Up-to-date	3.3	2.6	3.3
Average	3.1	2.6	2.5

Table 5. Average results of three AI tools based on three research fields

Article Type	**(1)**	**(2)**	**(3)**
ChatGPT (Open AI)	3.8	2.4	3.2
Gemini (Google)	2.4	2.6	3
Copilot (Microsoft)	2.2	2	3.4

Regarding the usefulness of the three tools based on the fields of research

ChatGPT provided the most accurate information in the field of global citizenship education while it performed the worst in the area of classroom equipment for physics. Conversely, Gemini excelled in delivering high-quality information on the standards for History assessment under the 2018 General Education Curriculum but scored the lowest in global citizenship education. Similarly, Copilot, like Gemini, performed best in providing information on the standards for History assessment in the 2018 General Education Curriculum but was weakest in the area of classroom equipment for physics.

Specifically, in the field of global citizenship education research, ChatGPT is considered the most suitable tool, with an average score of 3.8/5. The experts provided specific assessments for each criterion.

Criterion 1 Accuracy: "*ChatGPT (OpenAI) provided relatively accurate information. However, some inaccuracies remained such as the section on the general goals of the GCED program and the integration matrix of GCED content in the moral education curriculum. Both Gemini and Copilot provided incorrect information, particularly in sections on the goals of GCED in Vietnam, the general goals of the GCED program, and the integration matrix of GCED content in moral education*."

Criterion 3 Relevance: "*ChatGPT (OpenAI) answers are relatively aligned with the given questions and the original article. However, some content is not fully appropriate, such as the general goals of the GCED program and the integration matrix of GCED content. Gemini's (Google) and Copilot's answers lack focus, often presenting unrelated content*."

Criterion 4 Logicality and Clarity*:* "*ChatGPT (OpenAI) answers are logically presented in response to the given questions and the original article, while Gemini's (Google) answers are often overly verbose, disorganized, and some content appears arbitrary and unnecessary. For instance, the explanation of issues in global citizenship education (GCED) was speculative; or in the section on the goals of GCED in Vietnam, it presented specific expressions of global citizenship in Vietnam; similarly, the section on GCED in moral education also included educational methods and approaches. Copilot's (Microsoft) responses are quite disjointed, even though they include citations. For example, the introduction section included not only the goals of GCED but also discussions on methods, organizational forms, and evaluation criteria*."

In the field of research on classroom equipment for physics, Gemini was considered the most suitable tool, with an average score of 2.6/5. The experts provided specific assessments for each criterion:

Criterion 1 Accuracy: "*ChatGPT received a score of only 1/5, due to providing inaccurate and scientifically flawed information. For example, it incorrectly described Circular 32/2018/TT-BGDĐT as a regulation on the construction and organization of teaching activities in the 2018 2018 General Curriculum Framework, while it actually issued the curriculum itself. Similarly, it misattributed Circular 32/2020/TT-BGDĐT as related to the 2018 curriculum, whereas it pertains to the regulations for middle and high schools. Additionally, while ChatGPT provided a comprehensive list of requirements for physics laboratory rooms, the accuracy was low, particularly regarding calculations and formulas, which lacked a solid scientific basis. Copilot received the same score of 1/5 for similar reasons, such as misidentifying the regulations on laboratory room requirements.*"

Criterion 2 Comprehensiveness: "*Gemini achieved the highest score of 3/5. Its answers are generally comprehensive and cover most aspects of the issues, with reasonable expansion. In contrast, ChatGPT and Copilot*

scored only 2/5, missing key elements. For example, ChatGPT provided only one circular related to the article, but it misidentified it as Circular 12/2020/TT-BGDĐT instead of 32/2020/TT-BGDĐT. Similarly, Copilot's response lacked essential information regarding laboratory equipment setups."

Criterion 5 Up-to-date: "*Both ChatGPT and Gemini scored 2/5, as their information was largely outdated, with only a few updates reflecting recent changes. For instance, the formula for determining the number of physics laboratories was not up to date. Copilot, on the other hand, provided relatively current information, including infrastructure recommendations for physics laboratories, and was the only tool to mention details like ceiling height, flooring, and specialized furniture*."

In the field of research on establishing history assessment standards, Copilot was considered the most suitable tool, with an average score of 3.4/5. The experts provided detailed assessments for each criterion:

Criterion 1 Accuracy: "*ChatGPT and Gemini both provide generally accurate information, though lacking depth in certain areas, particularly when it comes to defining and explaining competencies. All three tools - ChatGPT, Gemini, and Copilot - have some inaccuracies when describing historical competencies. For example, Copilot's detailed example in Appendix 12 on building assessment standards for grade 10 history students did not align with the correct methodology for constructing standards."*

Criterion 3 Relevance*:* "*ChatGPT, Gemini, and Copilot all provide relevant information to the questions but lack sufficient detail in certain areas. For example, ChatGPT appropriately addressed questions about standards, assessment standards, and the characteristics of history as a subject, but the responses could have been more comprehensive.*"

Criterion 4 Logic and Clarity: "*ChatGPT provides information that is logical and clear at a basic level, but it occasionally lacks coherence. For example, in Appendix 7 on connecting required achievements to historical competencies, ChatGPT listed four components of historical competency, whereas there are only three, splitting the second component into two. Additionally, in the introduction, it reversed the order of Circular 32/2018/TT-BGDĐT and Resolution 29-NQ/TW, which is inaccurate. Gemini's answers were deemed logical and clear, though lacking detail. Copilot scored similarly with logical and clear information but faltered in providing precise definitions for standards and types of standards relevant to educational contexts.*"

Regarding the usefulness of the three tools based on the criterion scores

Among the evaluation criteria, accuracy was rated the lowest for all three tools, with the scores of 2.6/5, 2.3/5, and 2/5, respectively. Despite this low accuracy, the relevance, logical and clarity, and up-to-date of the information were rated relatively high. Gemini was particularly noted for its logical structure and clarity, receiving a score of 3/5, while Copilot stood out for its up-to-date information, with a score of 3.3/5.

Regarding the usefulness of the three tools based on the average criteria scores

ChatGPT achieved the highest average score across the five criteria, with an overall score of 3.1/5, outperforming the other two tools in all sub-criteria. The relevance, logical structure, and currency of its information were consistently rated at 3.3/5 across all three educational science domains. Gemini and Copilot showed similar average scores of 2.6 and 2.5, respectively, with Copilot showing a particularly high score for the criterion of up-to-date information.

4.2. Discussion

The findings of this study are generally consistent with existing literature on the application of generative AI in academic research. As previous studies have indicated, AI tools like ChatGPT and Gemini can effectively support idea generation, report drafting, and literature synthesis (Rahman et al., 2023; Nguyen-Trung et al., 2023). This aligns with our results, where these tools were rated highest on criteria such as relevance, logicality, and clarity. However, as emphasized in the literature, accuracy and factual reliability remain persistent limitations (Van Dis et al., 2023; Xames & Shefa, 2023). This is also confirmed in our evaluation, where all three tools scored the lowest in terms of accuracy. In the Physics classroom equipment topic, for example, both ChatGPT and Copilot misattributed or incorrectly interpreted key regulations, leading to severe misinformation. This reflects the known problem of AI "hallucinations," in which systems produce plausible-sounding but factually incorrect outputs (Morgan, 2023). Moreover, the variation in performance across different educational domains reinforces the argument that while generative AI can be a powerful support tool, its effectiveness depends heavily on context and task specificity. These

findings further highlight the need for localized guidelines and cautious implementation strategies, particularly in settings like Vietnam's education sector, where curriculum standards and policy frameworks are highly specific.

5. CONCLUSIONS

This study provides a comparative evaluation of three AI-powered tools - ChatGPT, Gemini, and Copilot - highlighting their respective strengths in different research areas. The findings show that each tool excels in specific fields: ChatGPT delivered the most accurate information in global citizenship education, while Gemini performed best in providing high-quality data related to the standards for History assessment under the 2018 General Education Curriculum. Similarly, Copilot stood out in the same area but struggled with accuracy when applied to the domain of classroom equipment for physics. Across all three tools, ChatGPT achieved the highest average score across various criteria, while Copilot ranked the lowest in most categories, except for the criterion of up-to-date information and references.

Notably, among the evaluation criteria, accuracy scored the lowest across all tools, signaling a significant area for improvement in AI's ability to provide precise information. Conversely, the tools received relatively high ratings for relevance, logical clarity, and the currency of the data presented. This suggests that while AI systems like ChatGPT excel at structuring responses and generating coherent narratives, their factual correctness remains an issue. Therefore, AI tools should be leveraged with care, particularly in tasks that do not require absolute accuracy, such as idea generation, outlining, and providing structured arguments. When used with proper verification of information, these tools can significantly enhance academic work.

Regarding the future application and development of AI in academic research, the use of well-structured command frameworks such as: MYTV (M in "Mệnh lệnh" - the main command; Y in "Yêu cầu" - specific requirements; T in "Tiêu chí" - criteria for evaluating responses; V in "Ví dụ" - additional examples or explanations) is recommended. This approach can help users to more effectively direct AI tools in generating useful and relevant content. While ChatGPT and other similar tools are still in development, ongoing improvements will lead to increasingly accurate outputs. As AI systems evolve, the role of researchers is expected to shift. Rather than focusing heavily on sourcing materials, researchers will likely concentrate more on refining research ideas, conducting methodologies, and analyzing data to derive meaningful conclusions.

From an ethical standpoint, the use of generative AI in educational research raises important concerns. These include the potential for misinformation, the risk of over-reliance on AI-generated outputs, and issues of academic integrity and authorship (Peres et al., 2023; Xames & Shefa, 2023). Researchers must remain accountable for verifying information, ensuring proper attribution, and maintaining transparency about the role of AI in the research process.

In the long term, AI advancements are expected to gradually alter the research process and the role of scholars. The introduction of AI into academic workflows could streamline various stages of research, such as the introduction, literature review, hypothesis generation, and data analysis phases. AI's ability to quickly scan vast datasets, organize information logically, and generate preliminary analyses makes it an invaluable tool in the early stages of research, allowing researchers to focus on more critical tasks like experimental design and interpretation of results.

The practical and theoretical implications of this research are considerable. On a practical level, the study highlights how AI tools can be integrated into academic research to enhance productivity and efficiency. On a theoretical level, it opens discussions on the evolving nature of research methodologies and the increasing interdependence between human researchers and AI-generated content. As AI continues to improve, it is likely to play a more central role in shaping how academic research is conducted, ultimately transforming the scholarly landscape.

However, this study is not without limitations. First, the evaluation relied heavily on expert judgment based on a small sample of research topics, which may not fully represent the broader scope of educational research in Vietnam. While the selected topics - curriculum development, classroom facilities, and assessment standards - reflect key areas of the 2018 curriculum, the findings may differ if applied to other disciplines or educational levels. Second, the prompts given to AI tools were constructed in a controlled research environment and may not reflect the variability in real-world usage, where users' digital literacy and prompting skills vary significantly. Moreover, as the AI tools themselves are constantly evolving, the performance results captured in this study represent a specific moment in time and may change with future updates.

Conflict of Interest: No potential conflict of interest relevant to this article was reported.

REFERENCES

Alasker, A., Alsalamah, S., Alshathri, N., Almansour, N., Alsalamah, F., Alghafees, M., AlKhamees, M., & Alsaikhan, B. (2024). Performance of large language models (LLMs) in providing prostate cancer information. *BMC Urology*, *24*(1). https://doi.org/10.1186/s12894-024-01570-0

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445922

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020, May 28). *Language Models are Few-Shot Learners*. arXiv.org. https://arxiv.org/abs/2005.14165

Camillepack. (2024). *Microsoft 365 Copilot overview*. Microsoft Learn. https://learn.microsoft.com/en-us/copilot/microsoft-365/microsoft-365-copilot-overview

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8cd550d1

Gibson, D., Jackson, S., Shanmugasundaram, R., Seth, I., Siu, A., Ahmadi, N., Kam, J., Mehan, N., Thanigasalam, R., Jeffery, N., Patel, M. I., & Leslie, S. (2024). Evaluating the efficacy of ChatGPT as a patient education tool in prostate Cancer: A Multi-Metric Assessment (Preprint). *Journal of Medical Internet Research*, *26*, e55939. https://doi.org/10.2196/55939

Google. (2023). *Gemini: AI-powered tools for research*. https://research.google.com/gemini

Henson, R. (2015). Analysis of Variance (ANOVA). In *Elsevier eBooks* (pp. 477-481). https://doi.org/10.1016/b978-0-12-397025-1.00319-5

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75-105. https://doi.org/10.5555/2017212.2017217

Hochmair, H. H., Juhász, L., & Kemp, T. (2024). Correctness comparison of ChatGPT-4, Gemini, Claude-3, and Copilot for spatial tasks. *Transactions in GIS*. https://doi.org/10.1111/tgis.13233

Hosseini, M., Rasmussen, L. M., & Resnik, D. B. (2023). Using AI to write scholarly publications. *Accountability in Research*, 1-9. https://doi.org/10.1080/08989621.2023.2168535

Kaftan, A. N., Hussain, M. K., & Naser, F. H. (2024). Response accuracy of ChatGPT 3.5 Copilot and Gemini in interpreting biochemical laboratory data: a pilot study. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-58964-1

KelliDavis. (2024). *Overview of Copilot*. Microsoft Learn. https://learn.microsoft.com/en-us/copilot/overview

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. https://doi.org/10.1371/journal.pdig.0000198

Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education*, *21*(2), 100790. https://doi.org/10.1016/j.ijme.2023.100790

Megahed, F. M., Chen, Y., Ferris, J. A., Knoth, S., & Jones-Farmer, L. A. (2023). How generative AI models such as ChatGPT can be (mis)used in SPC practice, education, and research? An exploratory study. *Quality Engineering*, 1-29. https://doi.org/10.1080/08982112.2023.2206479

Microsoft. (2023). *Copilot: AI for knowledge workers*. https://www.microsoft.com/copilot

MOET (Ministry of Education and Training). (2018). *Vietnam's 2018 new curriculum framework.* Ministry of Education and Training, Vietnam.

Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: the case of ChatGPT. *International Journal of Qualitative Methods*, *22*. https://doi.org/10.1177/16094069231211248

Nguyen-Trung, K. (2024). ChatGPT in Thematic Analysis: Can AI become a research assistant in qualitative Research? *OSF Preprints*. In press. https://doi.org/10.31219/osf.io/vefwc

Nguyen-Trung, K., Saeri, A. K., & Kaufman, S. (2023). Applying ChatGPT and AI-powered tools to accelerate evidence reviews. *OSF*. In press. https://doi.org/10.31219/osf.io/pcrqf

Nikolic, S., Sandison, C., Haque, R., Daniel, S., Grundy, S., Belkina, M., Lyden, S., Hassan, G. M., & Neal, P. (2024). ChatGPT, Copilot, Gemini, SciSpace and Wolfram versus higher education assessments: an updated multi-institutional study of the academic integrity impacts of Generative Artificial Intelligence (GenAI) on assessment, teaching and learning in engineering. *Australasian Journal of Engineering Education*, 1-28. https://doi.org/10.1080/22054952.2024.2372154

OpenAI (2023). *ChatGPT: AI models for advanced language tasks*. https://openai.com/chatgpt

Peres, R., Schreier, M., Schweidel, D., & Sorescu, A. (2023). On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. *International Journal of Research in Marketing*, *40*(2), 269-275. https://doi.org/10.1016/j.ijresmar.2023.03.001

Rahman, M., Terano, H. J. R., Rahman, N., Salamzadeh, A., & Rahaman, S. (2023). ChatGPT and Academic Research: A review and recommendations based on practical examples. *Journal of Education Management and Development Studies*, *3*(1), 1-12. https://doi.org/10.52631/jemds.v3i1.175

Rane, N., Choudhary, S., & Rane, J. (2024). Gemini versus ChatGPT: Applications, performance, architecture, capabilities, and implementation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4723687

Rossettini, G., Rodeghiero, L., Corradi, F., Cook, C., Pillastrini, P., Turolla, A., Castellini, G., Chiappinotto, S., Gianola, S., & Palese, A. (2024). Comparative accuracy of ChatGPT-4, Microsoft Copilot and Google Gemini in the Italian entrance test for healthcare sciences degrees: a cross-sectional study. *BMC Medical Education*, *24*(1). https://doi.org/10.1186/s12909-024-05630-9

Shang, L., Li, R., Xue, M., Guo, Q., & Hou, Y. (2024). Evaluating the application of ChatGPT in China's residency training education: An exploratory study. *Medical Teacher*, 1-7. https://doi.org/10.1080/0142159x.2024.2377808

Singh, H., & Singh, A. (2023). ChatGPT: Systematic Review, Applications, and Agenda for Multidisciplinary Research. *Journal of Chinese Economic and Business Studies*, *21*(2), 193-212. https://doi.org/10.1080/14765284.2023.2210482

Songkram, N., Chootongchai, S., Keereerat, C., & Songkram, N. (2024). Potential of ChatGPT in academic research: exploring innovative thinking skills. *Interactive Learning Environments*, 1-23. https://doi.org/10.1080/10494820.2024.2375342

UNESCO (2016). *Education for people and planet: Creating sustainable futures for all.* Global Education Monitoring Report.

Van Dis, E. a. M., Bollen, J., Zuidema, W., Van Rooij, R., & Bockting, C. L. (2023). ChatGPT: five priorities for research. *Nature*, *614*(7947), 224-226. https://doi.org/10.1038/d41586-023-00288-7

Yenduri, G., M, R., G, C. S., Y, S., Srivastava, G., Maddikunta, P. K. R., G, D. R., Jhaveri, R. H., B, P., Wang, W., Vasilakos, A., V., & Gadekallu, T. R. (2023). *Generative Pre-trained Transformer: a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.10435

Yıldız, M. S. (2023). Comparing response performances of ChatGPT-3.5, ChatGPT-4 and Bard to Health-Related questions: comprehensiveness, accuracy and being Up-to-Date. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4503443

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). Sage Publications.