

Evaluating Generative AI Tools for Student Learning: An Analytic Hierarchy Process Approach

Nikith Nannapaneni¹, Dr. A.V.S. Prasad²

¹The Global Edge School, Hyderabad, Telangana, India
Email: [nikithnannapaneni2026\[at\]gmail.com](mailto:nikithnannapaneni2026[at]gmail.com)

²SR Gayathri College, Hyderabad, Telangana, India
Email: [prasadavsp\[at\]gmail.com](mailto:prasadavsp[at]gmail.com)

Abstract: *This study evaluates seven generative AI platforms for student learning using the Analytic Hierarchy Process. Seven criteria were considered: accuracy, explanation quality, mathematical ability, research assistance, coding support, ease of use, and reliability. Pairwise-comparison data were collected from 500 students and analyzed to derive criterion and alternative priorities. The reported model assigns the greatest importance to research assistance and explanation quality. The composite evaluation ranks ChatGPT first, followed by Gemini, Qwen, Perplexity Claude, Microsoft Copilot, and DeepSeek. The comparison indicates that platform suitability varies across academic tasks and that overall rankings should be interpreted according to the criteria and judgments used in the AHP model. The findings provide a structured framework for selecting generative AI tools for different student-learning requirements.*

Keywords: Analytic Hierarchy Process (AHP), ChatGPT, Gemini, Claude, Perplexity, Microsoft Copilot, DeepSeek, and Qwen, Generative artificial intelligence, Multi-criteria decision-making, and Educational technology

1. Introduction

Generative AI platforms are increasingly used in education to support knowledge acquisition and complex learning tasks (Díaz, 2025; Seco et al., 2025). However, their capabilities vary, so their usefulness should be evaluated across the academic tasks students undertake (Xia et al., 2024). Although these tools can support brainstorming and writing feedback, they may also produce inaccurate or biased content, making critical use essential. This study compares seven generative AI platforms ChatGPT, Gemini, Claude, Perplexity, Microsoft Copilot, DeepSeek, and Qwen using the Analytic Hierarchy Process. It evaluates the tools across criteria including Accuracy, Explanation quality, Mathematical ability, Research assistance, Coding support and Ease of use to provide a structured comparison of their suitability for student learning (Akargöl et al., 2024; Veljić et al., 2025). By employing the Analytic Hierarchy Process, this research addresses the need for a robust decision-support framework that captures the interrelationships among these multifaceted technological criteria (Zeng et al., 2026). This systematic evaluation provides students with an objective methodology to navigate the diverse landscape of AI assistance, ensuring that platform selection is aligned with specific academic requirements (Đukić et al., 2024; Liu et al., 2025). These seven alternatives offer distinct forms of support for student learning, and their features may vary by model version, subscription, settings, and institutional access. The seven alternatives are elaborated below.

- **ChatGPT** provides conversational assistance for explaining concepts, answering questions, brainstorming, summarizing readings, revising text, solving problems, and generating or explaining code. Students can ask follow-up questions to clarify or adapt responses, but should check factual claims, calculations, and references.
- **Gemini** supports question answering, writing, summarization, and learning activities. Depending on the version and available integrations, it may also work with different types of input and connect with Google services.

Students should confirm that its responses and any cited information are accurate and suitable for their task.

- **Claude** offers conversational support for explaining complex material, developing and organizing ideas, summarizing longer texts, revising writing, and assisting with some coding and problem-solving tasks. Its usefulness depends on the task and the information provided; students should review its reasoning and verify important claims.
- **Perplexity** combines conversational answers with web search and links to sources. This can help students explore a topic, identify material for further reading, and begin research. Links do not guarantee that a source is authoritative or that the answer represents it accurately, so students should consult and evaluate the original sources.
- **Microsoft Copilot** can assist with question answering, drafting and revising text, summarizing information, problem-solving, and coding. Its available functions may depend on the specific Copilot product, user settings, and integration with Microsoft services. Students should review its outputs before relying on them in academic work.
- **DeepSeek** provides conversational assistance for tasks such as answering questions, explaining concepts, generating and revising text, and supporting mathematical or coding problems. Capabilities differ across models and versions, so students should test responses against course materials and independently check calculations, code, and factual claims.
- **Qwen** is a sophisticated suite of large language models and multimodal artificial intelligence systems developed by Alibaba's Tongyi Laboratory, represents a cutting-edge advancement in natural language processing and computational intelligence. This comprehensive family of models demonstrates exceptional capabilities across text generation, software development, complex analytical reasoning, and multimedia content processing.

Volume 15 Issue 10, October 2026

Fully Refereed | Open Access | Double Blind Peer Reviewed Journal

www.ijsr.net

Together, these tools may support a range of academic activities, but none should be treated as an authoritative source. The comparison considers seven criteria, each representing an aspect of a generative AI tool's usefulness for student learning. The criteria are elaborated below.

- **Accuracy** refers to the factual correctness, relevance, and task-specific appropriateness of a tool's responses, including explanations, calculations, code, and research-related information. It concerns the quality of an individual response, not consistency over time.
- **Explanation quality** concerns how clearly and coherently a tool communicates information and reasoning. Strong explanations present steps logically, define necessary terms, and suit the student's question and level of understanding.
- **Mathematical ability** evaluates a tool's capacity to interpret problems, apply appropriate methods, calculate correctly, and explain its solution process, including when symbolic expressions or multi-step reasoning are involved.
- **Research assistance** measures how effectively a tool helps students locate, evaluate, synthesize, and organize information. When it provides sources or references, their existence, relevance, and accuracy are part of this criterion; students should verify them independently.
- **Coding support** concerns a tool's ability to generate, explain, debug, and revise code, including whether its suggestions are appropriate to the requested programming language, tools, and constraints.
- **Ease of use** captures how accessible and straightforward a tool is in academic contexts, including the clarity of its interface, the effort required to interact with it, and the availability of useful functions.
- **Reliability** refers to the consistency of a tool's performance across similar tasks and repeated interactions. Unlike accuracy, which concerns whether a particular response is correct, reliability concerns how dependably the tool performs over time.

AHP provides a structured way to compare alternatives across multiple criteria, assess their relative importance, and make trade-offs explicit (Mohammed & Daham, 2021). This multidimensional approach is important because generative AI responses can be inaccurate or misleading, requiring students to assess outputs critically.

2. Objective of The Study

The study applies the Analytic Hierarchy Process to compare the platforms systematically across these criteria and account for their relative importance. It aims to identify their comparative strengths and limitations and clarify how their suitability may vary by academic task and learning need, rather than assuming that one platform will be best in every area. Although prior studies have examined AI assistants for academic research and educational applications, student-centred comparisons across specific academic tasks and learning-related criteria remain underdeveloped (Akin et al., 2025; Isiaku et al., 2024). By evaluating these tools, this research addresses the need to understand how AI agents may support student learning, self-regulation, and task performance (Fan et al., 2024). The comparison considers multiple dimensions because students use generative AI for

varied academic purposes, including tailored learning support and research assistance. The literature review that follows situates this comparison within research on students' uses and perceptions of generative AI and prior multi-criteria evaluations of AI tools in education.

3. Literature Review

Students use generative AI for more than assignment support: reported academic applications include clarifying concepts, obtaining tailored feedback, brainstorming, summarizing readings, and developing initial research ideas, reflecting expectations that these tools can support both independent learning and academic productivity. These benefits coexist with concerns about inaccurate or biased outputs and unreliable references, underscoring the need for students to verify generated material rather than treat it as authoritative (Chan & Hu, 2023). Prior multi-criteria studies have applied AHP and related methods to educational AI questions, including evaluating digital learning resources and considering issues such as usability, trust, academic integrity, and technical characteristics (Rădulescu & Rădulescu, 2025). Other selection frameworks emphasize accuracy, relevance, comprehensibility, stability, user-friendliness, and cost, illustrating that judgments of tool suitability depend on multiple dimensions rather than a single measure of capability. ChatGPT, Gemini, Claude, Perplexity, Microsoft Copilot, DeepSeek, and Qwen. Using student-centred criteria that address response quality, learning support, academic task performance, usability, and dependability. By applying AHP to these dimensions, the study provides a structured comparison of tools whose relative strengths may differ across the needs and tasks of students.

This methodology facilitates a nuanced understanding of how each platform aligns with specific pedagogical requirements, such as the need for precise research assistance or robust coding support (Ghorbani et al., 2025). Such granular evaluation allows educators and students to move beyond anecdotal tool preference, establishing a rigorous framework for identifying platforms that best balance computational power with the inherent risks of generative output. Furthermore, by quantifying the comparative weight of these factors, the model illuminates the significant variance in how platforms prioritize search-based verification versus generative creativity. Consequently, this study integrates these diverse performance metrics to offer a holistic assessment of contemporary AI systems. This systematic approach ultimately provides actionable insights into selecting generative AI resources that align with individual learning objectives and institutional academic standards.

AHP has been applied to educational technology selection by structuring decisions into criteria and alternatives, using pairwise comparisons to derive priorities; for example, it has been used to select engineering education software and compare AI tools for university teaching and learning (De-La-Cruz-Vasquez et al., 2024; Dorado-Vicente et al., 2011). Related studies extend this approach to generative AI adoption: a fuzzy AHP framework prioritizes concerns including academic integrity, biased outputs, and the erosion of critical thinking, while other reviews document AHP applications addressing usability, trust, educational-resource

quality, and ethical risks. Recent multi-criteria comparisons of academic language-model assistants further show the value of evaluating platforms across performance, user experience, ethical issues, and technical constraints, although they employ methods other than AHP (Akın et al., 2025). Integrating these multifaceted evaluation strategies, the present research employs AHP to systematically reconcile these competing performance dimensions, ensuring a balanced appraisal of generative AI in the context of academic rigor (Batista et al., 2024; Lin et al., 2025). Although prior AHP studies have selected educational software and AI tools for university teaching, and other multi-criteria comparisons have assessed academic assistants across performance, user experience, ethical, and technical dimensions, these approaches provide a basis for a more explicitly student-centred comparison.

The following methodology describes how the evaluation hierarchy is structured and how pairwise comparisons are used to derive the priorities and rankings. This hierarchy organizes seven distinct criteria accuracy, explanation quality, mathematical ability, research assistance, coding support, ease of use, reliability into a unified decision matrix.

4. Methodology

The Analytic Hierarchy Process (AHP) is one of the Multi-criteria-decision making methods(MCDM) and the decision-support framework developed by Saaty.T. L1980. Its main aim is to assess the relative importance of a defined set of alternatives using a ratio scale, which is based on the decision-maker's judgment. This methodology highlights the importance of the intuitive assessments made by the decision-maker and the need for consistency when comparing alternatives during the decision-making process. Since decision-makers depend on their expertise and experience to make judgments and ultimately decisions, the AHP framework is well-suited to their behavioural tendencies. A significant benefit of this method is its capacity to

systematically arrange both measurable and non-measurable factors, providing a structured yet relatively simple approach to addressing decision-making issues. Additionally, by logically breaking down a problem from a broader viewpoint to more specific details, one can create links between the smaller components and the larger context through straightforward paired comparison judgments. (Saaty.TL 1980) outlined the following steps for implementing the AHP:

- 1) Define the problem and determine its goal.
- 2) Structure the hierarchy from the top (the objectives from a decision-maker's viewpoint) through the intermediate levels (criteria on which sub-sequent levels depend) to the lowest level which usually contains the list of alternatives.
- 3) Construct a set of pair-wise comparison matrices (size $n \times n$) for each of the lower levels with one matrix for each element in the level immediately above by using the relative scale measurement shown in Table 1. The pair-wise comparisons are done in terms of which element dominates the other.
- 4) There are $n(n-1)/2$ judgments required to develop the set of matrices in step 3. Reciprocals are automatically assigned in each pair-wise comparison.
- 5) Hierarchical synthesis is now used to weight the eigenvectors by the weights of the criteria and the sum is taken over all weighted eigenvector entries corresponding to those in the next lower level of the hierarchy.
- 6) Having made all the pair-wise comparisons, the consistency is determined by using the eigenvalue, $\lambda_{\max}$, to calculate the consistency index, CI as follows: $CI = (\lambda_{\max} - n) / (n - 1)$, where n is the matrix size. Judgment consistency can be checked by taking the consistency ratio (CR) of CI with the appropriate value in Table 2. The CR is acceptable, if it does not exceed 0.10. If it is more, the judgment matrix is inconsistent. To obtain a consistent matrix, judgments should be reviewed and improved.
- 7) Steps 3-6 are performed for all levels in the hierarchy.

Table 1: The saaty's scale of relative importance

Relative importance	Definition	Description
1	Equal importance	Two factors equally influence the objective
3	Moderate importance	Experience and judgement slightly favour one factor over another
5	Strong importance	Experience and judgement strongly favour one factor over another
7	Very strong importance	One decision factor is strongly favoured over another, and its supremacy is established in practice
9	Extreme importance	The evidence favouring one decision factor over another is of the highest possible order of validity
2,4,6 and 8	Intermediate values between adjacent values	When compromise is required

Table 2: Average random consistency (RI)

Size of matrix	1	2	3	4	5	6	7	8	9	10
Random consistency	0	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49

A systematically organized questionnaire was utilized to conduct the data collection. The questionnaire was carefully constructed to maximize the precision of the information obtained and to improve the clarity of questions for respondents. Data were collected from 500 students within the age range of 19 to 28 years of different regions in India.

Following data collection, the Analytic Hierarchy Process (AHP) methodology was employed to analyze the gathered data and achieve the research objectives. The figure -1 shows the proposed model and AHP calculations are given in Table 3 to 11.

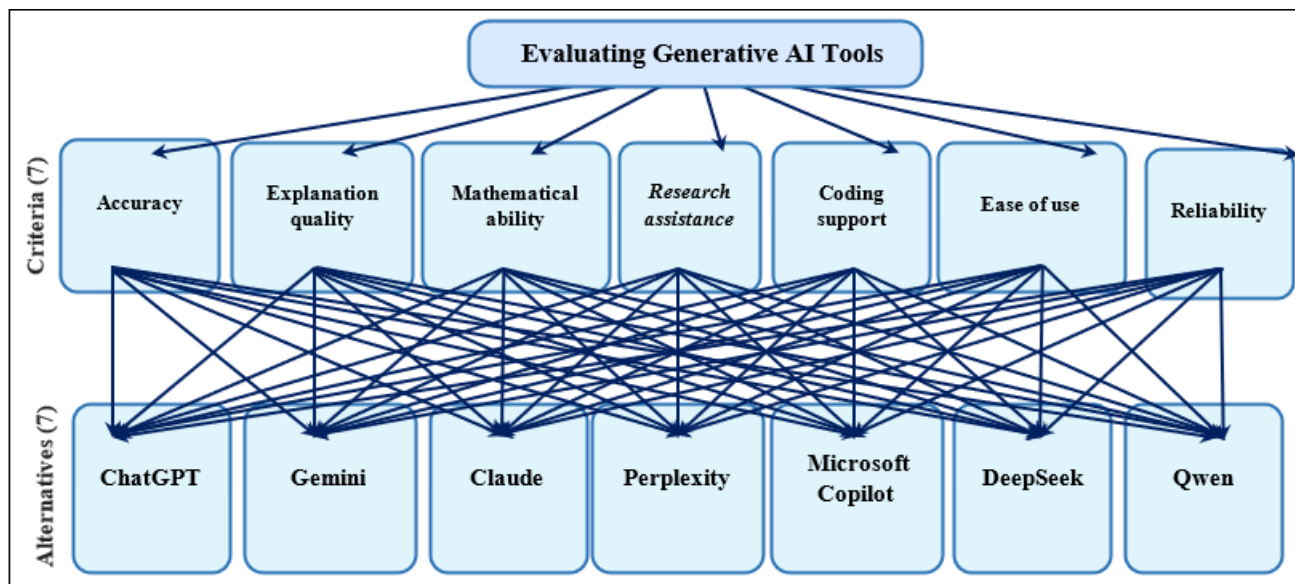


Figure 1: AHP Hierarchical structure with 7 criteria and 7 Alternatives

Table 3: Pair-wise comparison matrix for all criteria

	Accuracy	Explanation quality	Mathematical ability	Research assistance	Coding support	Ease of use	Reliability	Priority vector
Accuracy	1	1/2	2	1/3	1/5	3	1/3	0.103
Explanation quality	2	1	2	1/2	4	3	2	0.209
Mathematical ability	1/2	1/2	1	1/2	2	2	2	0.131
Research assistance	3	2	2	1	2	2	2	0.233
Coding support	5	1/4	1/2	1/2	1	3	1/2	0.129
Ease of use	1/3	1/3	1/2	1/2	1/3	1	1/3	0.058
Reliability	3	1/2	1/2	1/2	2	3	1	0.137
$\lambda_{\max} = 7.763$, CI=0.127, CR = 0.096 < 0.1								

Table 4: Pairwise comparison matrix for “Accuracy “

	Perplexity	Claude	Gemini	ChatGPT	Microsoft Copilot	Qwen	DeepSeek	Priority vector
Perplexity	1	3	2	3	1/2	1/4	1/5	0.163
Claude	1/3	1	1/3	1/5	3	1/3	3	0.103
Gemini	1/2	3	1	1/4	4	2	2	0.174
ChatGPT	1/3	5	4	1	2	1/2	3	0.210
Microsoft Copilot	2	1/3	1/4	1/2	1	1/3	2	0.119
Qwen	4	3	1/2	2	3	1	1/2	0.144
DeepSeek	5	1/3	1/2	1/3	1/2	2	1	0.087
$\lambda_{\max} = 7.588$, CI=0.098 CR = 0.07 < 0.1								

Table 5: Pairwise comparison matrix for “Explanation quality”

	Perplexity	Claude	Gemini	ChatGPT	Microsoft Copilot	Qwen	DeepSeek	Priority vector
Perplexity	1	4	2	2	1/2	1/4	4	0.179
Claude	1/4	1	1/3	1/4	1/3	1/2	3	0.067
Gemini	1/2	3	1	1/4	4	2	2	0.169
ChatGPT	1/2	4	4	1	2	1/2	3	0.195
Microsoft Copilot	2	3	1/4	1/2	1	1/3	2	0.117
Qwen	4	2	1/2	2	3	1	1/2	0.191
DeepSeek	1/4	1/3	1/2	1/3	1/2	2	1	0.082
$\lambda_{\max} = 7.679$, CI=0.113 CR = 0.09 < 0.1								

Table-6: Pair-wise comparison matrix for “Mathematical ability”

	Perplexity	Claude	Gemini	ChatGPT	Microsoft Copilot	Qwen	DeepSeek	Priority vector
Perplexity	1	2	1/4	1/2	3	1/4	4	0.133
Claude	1/2	1	1/3	1/4	3	3	3	0.140
Gemini	4	3	1	1/4	2	2	2	0.204
ChatGPT	2	4	4	1	2	1/2	3	0.251
Microsoft Copilot	1/3	1/3	1/2	1/2	1	1/3	2	0.068
Qwen	4	1/3	1/2	2	3	1	1/2	0.129
DeepSeek	1/4	1/3	1/2	1/3	1/2	2	1	0.074
$\lambda_{\max} = 7.423$, CI=0.071, CR = 0.05 < 0.1								

Table 7: Pair-wise comparison matrix for “*Research assistance*”

	Perplexity	Claude	Gemini	ChatGPT	Microsoft Copilot	Qwen	DeepSeek	Priority vector
Perplexity	1	2	1/4	1/2	3	1/4	1/2	0.115
Claude	1/2	1	1/3	1/4	3	4	4	0.146
Gemini	4	3	1	1/4	2	6	2	0.236
ChatGPT	2	4	4	1	2	1/2	3	0.255
Microsoft Copilot	1/3	1/3	1/2	1/2	1	1/3	2	0.068
Qwen	4	1/4	1/6	2	3	1	1/2	0.116
DeepSeek	2	1/4	1/2	1/3	1/2	2	1	0.063
$\lambda_{\max} = 7.633$, CI=0.105, CR = 0.080 < 0.1								

Table 8: Pair-wise comparison matrix for “*coding support*”

	Perplexity	Claude	Gemini	ChatGPT	Microsoft Copilot	Qwen	DeepSeek	Priority vector
Perplexity	1	3	1/4	1/2	3	1/4	2	0.127
Claude	1/3	1	1/3	1/4	3	5	3	0.153
Gemini	4	3	1	1/4	2	3	2	0.210
ChatGPT	2	4	4	1	2	1/2	3	0.252
Microsoft Copilot	1/3	1/3	1/2	1/2	1	1/3	2	0.070
Qwen	4	1/5	1/3	2	3	1	1/2	0.121
DeepSeek	1/2	1/3	1/2	1/3	1/2	2	1	0.068
$\lambda_{\max} = 7.656$, CI=0.109, CR = 0.083 < 0.1								

Table 9: Pair-wise comparison matrix for “*Ease of use*”

	Perplexity	Claude	Gemini	ChatGPT	Microsoft Copilot	Qwen	DeepSeek	Priority vector
Perplexity	1	4	1/4	1/2	3	1/4	2	0.147
Claude	1/4	1	1/3	2	3	5	3	0.193
Gemini	4	3	1	1/4	2	2	3	0.217
ChatGPT	2	1/2	4	1	2	1/2	3	0.203
Microsoft Copilot	1/3	1/3	1/2	1/2	1	1/3	2	0.066
Qwen	4	1/5	1/2	2	3	1	1/2	0.110
DeepSeek	1/2	1/3	1/3	1/3	1/2	2	1	0.064
$\lambda_{\max} = 7.727$, CI=0.121 CR = 0.092 < 0.1								

Table 10: Pair-wise comparison matrix for “*Reliability*”

	Perplexity	Claude	Gemini	ChatGPT	Microsoft Copilot	Qwen	DeepSeek	Priority vector
Perplexity	1	4	1/4	1/2	1/3	1/4	2	0.101
Claude	1/4	1	1/3	2	2	1/4	3	0.124
Gemini	4	3	1	1/4	3	2	3	0.222
ChatGPT	2	1/2	4	1	2	1/2	3	0.203
Microsoft Copilot	3	1/2	1/3	1/2	1	1/3	2	0.102
Qwen	4	4	1/2	2	3	1	1/2	0.167
DeepSeek	1/2	1/3	1/3	1/3	1/2	2	1	0.082
$\lambda_{\max} = 7.752$, CI=0.125, CR = 0.095 < 0.1								

Table 11: Final priority table

	Accuracy	Explanation quality	Mathematical ability	Research assistance	Coding support	Ease of use	Reliability	Final Priority vector	Ranking
	0.103	0.209	0.131	0.233	0.129	0.058	0.137		
Perplexity	0.163	0.179	0.133	0.115	0.127	0.147	0.101	0.1372	4
Claude	0.103	0.067	0.140	0.146	0.153	0.193	0.124	0.1249	5
Gemini	0.174	0.169	0.204	0.236	0.210	0.217	0.222	0.2050	2
ChatGPT	0.210	0.195	0.251	0.255	0.252	0.203	0.203	0.2267	1
Microsoft Copilot	0.119	0.117	0.068	0.068	0.070	0.066	0.102	0.0882	6
Qwen	0.144	0.191	0.129	0.116	0.121	0.110	0.167	0.1418	3
DeepSeek	0.087	0.082	0.074	0.063	0.068	0.064	0.082	0.0742	7

5. Results

The AHP model was applied to the pairwise-comparison data collected from 500 students aged 19 to 28 years. Priority vectors were derived for the seven criteria and, within each criterion, for the seven generative AI platforms; these were then synthesised into a composite priority vector for the alternatives. The consistency of every comparison matrix was examined using the consistency index and consistency ratio,

with a CR not exceeding 0.10 taken as acceptable. All matrices satisfied this condition (Tables 3–10), indicating that the judgments supplied by the respondents were sufficiently consistent to support the derived priorities.

Criterion priorities: Table 3 reports the pairwise comparison of the seven criteria and the resulting priority vector. Research assistance emerged as the most heavily weighted criterion (0.233), followed by explanation quality (0.209). Reliability

(0.137), mathematical ability (0.131) and coding support (0.129) formed a middle group of closely comparable importance. Accuracy received a weight of 0.103, and ease of use ranked last (0.058). The pattern indicates that respondents placed the greatest emphasis on a tool's capacity to help them locate, evaluate, synthesise and organise information, and on how clearly it communicates its reasoning, while interface simplicity and accessibility were comparatively less decisive in their judgments.

Alternative priorities by criterion. Tables 4–10 show the alternative priorities.

- **Accuracy:** ChatGPT obtained the highest priority (0.210), followed by Gemini (0.174) and Perplexity (0.163). Qwen (0.144) and Microsoft Copilot (0.119) followed, while Claude (0.103) and DeepSeek (0.087) received the lowest priorities.
- **Explanation quality:** ChatGPT again led (0.195), but only narrowly ahead of Qwen (0.191), with Perplexity (0.179) and Gemini (0.169) close behind. Microsoft Copilot (0.117), DeepSeek (0.082) and Claude (0.067) were ranked lower.
- **Mathematical ability:** ChatGPT (0.251) was clearly favoured, followed by Gemini (0.204). Claude (0.140), Perplexity (0.133) and Qwen (0.129) were closely grouped, while DeepSeek (0.074) and Microsoft Copilot (0.068) received the lowest weights.
- **Research assistance:** ChatGPT (0.255) and Gemini (0.236) were the two strongest, ahead of Claude (0.146), Qwen (0.116) and Perplexity (0.115). Microsoft Copilot (0.068) and DeepSeek (0.063) were ranked last.
- **Coding support:** ChatGPT (0.252) led, followed by Gemini (0.210) and Claude (0.153); Perplexity (0.127) and Qwen (0.121) were intermediate, and Microsoft Copilot (0.070) and DeepSeek (0.068) were lowest.
- **Ease of use:** the ordering changed, with Gemini (0.217) ranking first, followed closely by ChatGPT (0.203) and Claude (0.193). Perplexity (0.147) and Qwen (0.110) followed, while Microsoft Copilot (0.066) and DeepSeek (0.064) were ranked lowest.
- **Reliability:** Gemini (0.222) obtained the highest priority, ahead of ChatGPT (0.203) and Qwen (0.167). Claude (0.124), Microsoft Copilot (0.102) and Perplexity (0.101) were closely grouped, and DeepSeek (0.082) was ranked last.

Composite priorities and ranking. Table-11 synthesises the criterion weights with the alternative priorities to produce the final priority vector. ChatGPT ranked first with a composite priority of 0.2267, followed by Gemini (0.2050). Qwen (0.1418) and Perplexity (0.1372) occupied the middle of the ranking, with Claude (0.1249) close behind. Microsoft Copilot (0.0882) and DeepSeek (0.0742) were ranked sixth and seventh respectively.

Taken together, the composite results indicate that, under the criteria weights and judgments used in this model, ChatGPT and Gemini offer the most balanced overall support for student learning, combining strong performance on accuracy, explanation quality, mathematical ability, research assistance and coding support with competitive ease of use and reliability. Qwen, Perplexity and Claude form a middle tier whose standing varies by criterion, while Microsoft Copilot

and DeepSeek were judged comparatively less suitable across the criteria examined. Notably, the platforms that lead on some criteria do not lead on others. ChatGPT ranked first on the five performance-related criteria, whereas Gemini ranked first on ease of use and reliability. This divergence illustrates that tool suitability is task-dependent rather than a single overall measure of capability.

6. Conclusion

This study applied the Analytic Hierarchy Process to evaluate seven generative AI platforms ChatGPT, Gemini, Claude, Perplexity, Microsoft Copilot, DeepSeek and Qwen for student learning across seven criteria: accuracy, explanation quality, mathematical ability, research assistance, coding support, ease of use and reliability. Pairwise-comparison data from 500 students were analysed to derive criterion weights, per-criterion alternative priorities and a composite ranking.

The findings provide a structured student-centred decision framework for selecting among generative AI tools. Rather than relying on anecdotal preference, students and educators can use the criterion weights and composite priorities to match a platform to a particular requirement, for example favouring ChatGPT or Gemini for quantitative and research intensive work, or Gemini where ease of use and dependable everyday interaction are the priority. The explicit weighting also makes trade-offs visible, allowing a user who prioritises a different criterion to see how that emphasis reshapes the overall ordering.

Several limitations should be acknowledged. The results reflect the criteria and judgments supplied by the respondents in this study and the specific pairwise comparisons used, different weights or judgments would change the composite ranking. The comparison also treats each platform as a single alternative, whereas features and performance may vary by model version, subscription tier, settings and institutional access. The study evaluated perceived suitability for student learning rather than objective benchmark performance, and the criteria of accuracy and reliability refer respectively to the quality of an individual response and to consistency over time, both of which are difficult to capture fully through pairwise judgments. Finally, because generative AI tools are updated frequently, the ranking should be read as a snapshot of the platforms as assessed at the time of data collection.

Future research could replicate the model with different or larger respondent groups, compare results across disciplines and levels of study, and combine AHP with complementary multi-criteria methods to triangulate the rankings. Extending the framework to include newer platforms, updated model versions and additional criteria such as cost, data privacy and academic-integrity safeguards would further strengthen its usefulness for students and institutions.

References

- [1] Akargöl, İ., Karadağ, İ., & Gürçan, Ö. F. (2024). Selecting The Optimal E-Learning Platform for Universities: A Pythagorean Fuzzy AHP/TOPSIS Evaluation. *The European Journal of Research and*

- Development*, 4(2), 19–34. <https://doi.org/10.56038/ejrnd.v4i2.425>
- [2] Akın, M., Sir, G. D. B., Aydemir-Karadag, A., & Çerçioğlu, H. (2025). A Multi-Criteria Comparison of Large Language Model Powered Assistants in Pre-Research Studies for the Academia. *IEEE Access*, 13, 127086–127099. <https://doi.org/10.1109/access.2025.3586502>
- [3] Batista, J., Mesquita, A., & Carnaz, G. (2024). Generative AI and Higher Education: Trends, Challenges, and Future Directions from a Systematic Literature Review. *Information*, 15(11), 676. <https://doi.org/10.3390/info15110676>
- [4] Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20(1). <https://doi.org/10.1186/s41239-023-00411-8>
- [5] De-La-Cruz-Vasquez, A., Flor-Cunza, H., Lara-Herrera, J., & Romero-Untiveros, L. (2024). *Selection of the Best AI Educational Tool Applied to University-Level Teaching and Learning Process Using the AHP Methodology*. International Conference on Advanced Computing and Innovative Technologies in Engineering (ICACIT) 1–6. <https://doi.org/10.1109/icacit62963.2024.10788591>
- [6] Díaz, G. M. (2025). Supporting Reflective AI Use in Education: A Fuzzy-Explainable Model for Identifying Cognitive Risk Profiles. *Education Sciences*, 15(7), 923. <https://doi.org/10.3390/educsci15070923>
- [7] Dorado-Vicente, R., Gómez-Moreno, A., Torres-Jiménez, E., & López-Alba, E. (2011). An AHP application to select software for engineering education. *Computer Applications in Engineering Education*, 22(2), 200–208. <https://doi.org/10.1002/cae.20546>
- [8] Đukić, T., Novaković, S., & Jauković-Jocić, K. (2024). Evaluating the significance of artificial intelligence (AI) in academic platforms by using PIPRECIA-S method. *Ekonomika*, 70(3), 11–24. <https://doi.org/10.5937/ekonomika2403011d>
- [9] Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, T., Shen, Y., Li, X., & Gašević, D. (2024). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56(2), 489–530. <https://doi.org/10.1111/bjet.13544>
- [10] Ghorbani, M. K., Talebbeydokhti, N., Hamidifar, H., Samadi, M., Nones, M., Rezaeitavabe, F., & Heidarifar, S. (2025). Application of Multi-Criteria Decision-Making Models for Assessment of Education Quality in Water Resources Engineering. *Algorithms*, 18(1), 12. <https://doi.org/10.3390/a18010012>
- [11] Isiaku, L., Muhammad, A. S., Kefas, H. I., & Ukaegbu, F. C. (2024). Enhancing technological sustainability in academia: leveraging ChatGPT for teaching, learning and evaluation. *Quality Education for All*, 1(1), 385–416. <https://doi.org/10.1108/qea-07-2024-0055>
- [12] Lin, P., Deng, Q., & Zhou, Y. (2025). Towards responsible AI in education: A Delphi-AHP-based framework for evaluating educational large language models. *Computers and Education Artificial Intelligence*, 10, 100534. <https://doi.org/10.1016/j.caeai.2025.100534>
- [13] Liu, Y., Yang, X., Zhu, Y., Wang, J., Zuo, M., Yang, L., & Sun, L. (2025). Attribution-Driven Teaching Interventions: Linking I-AHP Weighted Assessment to Explainable Student Clustering. *Algorithms*, 18(11), 691. <https://doi.org/10.3390/a18110691>
- [14] Mohammed, H. J., & Daham, H. A. (2021). Analytic Hierarchy Process for Evaluating Flipped Classroom Learning. *Computers, Materials & Continua* 66(3), 2229–2239. <https://doi.org/10.32604/cmc.2021.014445>
- [15] Rădulescu, C. Z., & Rădulescu, M. (2025). Criteria Analysis for the Selection of a Generative Artificial Intelligence Tool for Academic Research Based on an Improved Group DEMATEL Method. *Applied Sciences*, 15(10), 5416. <https://doi.org/10.3390/app15105416>
- [16] Saaty TL.1980 The analytic hierarchy process. New York: McGraw-Hill.
- [17] Seco, D., Groesser, S. N., & Pedrosa, A. M. (2025). Use of Generative Artificial Intelligence tools in higher education environments. *Multidisciplinary Journal for Education Social and Technological Sciences*, 12(1), 156–175. <https://doi.org/10.4995/muse.2025.23623>
- [18] Veljić, A., Viduka, D., Ilić, L., Karabašević, D., Šijan, A., & Papic, M. (2025). Sustainable Decision-Making in Higher Education: An AHP-NWA Framework for Evaluating Learning Management Systems. *Sustainability*, 17(22), 10130. <https://doi.org/10.3390/su172210130>
- [19] Xia, Q., Weng, X., Ouyang, F., Lin, T., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education*, 21(1). <https://doi.org/10.1186/s41239-024-00468-z>
- [20] Zeng, Q., Zhang, H., & Li, C. (2026). Intelligent hybrid decision support systems for education policy and institutional performance optimization. *Scientific Reports*. <https://doi.org/10.1038/s41598-026-51394-1>.