

Generative artificial intelligence as an evaluator and feedback tool in distance learning: a case study on law implementation

Dian Nurdiana¹, Muhamad Riyan Maulana¹, Siti Hadijah Hasanah², Madiha Dzakiyyah Chairunnisa³,
Avelyn Pingkan Komuna³, Muhammad Rif'an¹

¹Information Systems Study Program, Faculty of Science and Technology, Universitas Terbuka, South Tangerang City, Indonesia

²Statistics Study Program, Faculty of Science and Technology, Universitas Terbuka, South Tangerang City, Indonesia

³Law Study Program, Faculty of Law, Social Sciences and Political Sciences, Universitas Terbuka, South Tangerang City, Indonesia

Article Info

Article history:

Received Jul 24, 2024

Revised Jan 28, 2025

Accepted Mar 15, 2025

Keywords:

AI evaluation

Assignment feedback

Distance learning

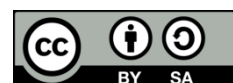
Generative artificial intelligence

Law science

ABSTRACT

The development of generative artificial intelligence (GAI) has impacted various fields, including higher education. This research examines the use of GAI as an evaluator and feedback provider in distance legal education. This study tested five GAI models: ChatGPT, Perplexity, Gemini, Bing, and You, using a sample of 20 students and evaluations from legal experts. Descriptive statistical analysis and non-parametric tests, including Wilcoxon, intraclass correlation coefficient (ICC), Kappa, and Kendall's W, were used to assess accuracy, feedback quality, and usability. The results showed that ChatGPT was the most effective GAI, with the highest mean scores of 4.22 from experts and 4.12 from students, followed by Gemini with scores of 4.15 and 4.07. In terms of binary judgement accuracy, Gemini scored 80%, ChatGPT 60%, while Perplexity, Bing, and You had lower scores. Statistical analysis showed moderate agreement (ICC=0.439) and low alignment (Kappa=-0.058) between the GAIs and expert evaluations, with a Kendall's W value of 0.576 indicating moderate consistency in judgements. These findings emphasize the importance of selecting effective GAIs such as ChatGPT and Gemini to improve academic evaluation and learning in legal education, and pave the way for further innovations in the use of AI.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Dian Nurdiana

Information Systems Study Program, Faculty of Science and Technology, Universitas Terbuka

St. Pd. Cabe Raya, Pd. Cabe Udik, Ciputat, South Tangerang City, Banten 15437, Indonesia

Email: dian.nurdiana@ecampus.ut.ac.id

1. INTRODUCTION

Generative development artificial intelligence (GAI) has now reached an impressive point, with the ability to perform increasingly complex tasks and approach human intelligence in various fields [1]. Advances in natural language processing technology, computer vision, and machine learning have enabled GAI to play a role in a variety of sectors, from healthcare, manufacturing, to educational services [2]–[4]. The benefits of GAI include increased operational efficiency, better decision-making based on faster and more accurate data analysis, and the ability to automate processes that require high-precision [5], [1]. In the education sector, GAI offers a variety of applications that can revolutionize the way teaching and learning are carried out [6], [7]. One example of the use of GAI in education is in the context of distance education, where GAI can help create a more interactive and personalized learning experience for students who are geographically separated from their instructors [8], [9].

In the context of distance education, GAI plays a significant role in improving the quality of interactions between teachers and students [10], [11]. GAI technology allows the creation of a more adaptive learning environment, where learning materials can be adapted to individual student needs and progress [12]. GAI can be used to provide study guides that are more structured and can be accessed at any time, helping students to remain involved in the learning process even though they are far from the instructor or lecturer [13], [14]. GAI-based virtual assistants can also support students in completing their assignments by providing necessary assistance and information [15]. Meanwhile, in the field of legal science, the implementation of GAI can contribute to helping students understand complex legal concepts and prepare them for in-depth legal case analysis [16].

Legal science is a complex and dynamic field, requiring an in-depth understanding of legal texts, juridical precedents, and critical analysis of various cases [17]–[19]. Evaluation in law focuses not only on theoretical understanding, but also on a student's ability to apply legal principles to factual situations [20], [21]. Feedback given to law students must be able to guide them in understanding the complexity of the law and developing the necessary analytical skills [22], [23]. The current use of GAI in legal science is limited to tools for searching for legal information, analyzing legal texts, and simulating simple cases [16]. Although this technology has been used for various purposes, its use in aspects of academic assessment still needs to be optimized [24]. Manually evaluating student assignments requires significant time and effort from instructors, which can reduce the time available for other teaching activities [25], [26]. By utilizing GAI, this process can be automated so that teachers can focus more on developing learning materials and interacting with students [27], [28].

In the context of distance education, GAI has excellent potential to overcome some of the main challenges in distance education [29]. GAI can allow for limited direct interaction between students and instructors, so timely and quality feedback is critical to maintaining student engagement and learning progress [30], [31]. This technology has the potential to provide fast, accurate, and personalized evaluations, which in turn can help students understand the material better and correct their mistakes more effectively [32], [33]. With the increasing adoption of distance education, mainly due to the COVID-19 pandemic, there is an urgent need for evaluation solutions that can address the challenges faced by teachers and students [34], [35].

Previous research has proven that AI can be used to personalize learning and data analysis to improve student learning outcomes [36]. The implications of previous research show that AI technology has excellent potential to increase the efficiency and effectiveness of the learning process. However, research specifically exploring the use of GAI in assignment evaluation and feedback is limited [37]. Meanwhile, Su and Yang [38] examined the use of ChatGPT in education through the IDEE framework, which includes desired results, level of automation, ethics, and evaluation of effectiveness. This research shows that ChatGPT can increase personalization and learning efficiency and improve teacher feedback. However, challenges include untested effectiveness, data quality, and ethical and safety issues. This study highlights the great potential of ChatGPT in education, but still emphasizes the need to overcome the challenges. Therefore, this research aims to fill this gap by exploring the possibility of GAI in the context of student assignment evaluation, especially in implementing laws in legal case studies.

The focus of this research is to explore and analyze the use of GAI as an evaluator and provider of feedback in student assignments in the distance education sector, focusing on legal case studies. This research methodology involves testing five types of GAI, namely ChatGPT, Perplexity, Gemini, Bing, and You, involving 20 students as samples. Each GAI will assess and provide feedback on assignments regarding the implementation of laws in legal case studies, which are then compared with the assessments and input from legal experts. The measurement variables include three main aspects: accuracy, quality of feedback, and usefulness of feedback for students.

Accuracy in this context refers to the extent to which GAI can provide assessments that comply with applicable academic and legal standards [39], [40]. Some literature states that the accuracy of AI in task evaluation depends on the algorithm used and the data on which it was trained [41]. This statement is in line with the research results of [42] who expressed the opinion that AI has great potential to provide accurate evaluations if it is trained with appropriate and relevant data.

Meanwhile, feedback quality involves evaluating how in-depth and valuable the feedback provided by GAI is [39]. Quality feedback not only points out errors, but also provides explanations that help students understand the correct concepts [43], [44]. According to Lipnevich and Panadero [45], in their study on educational feedback emphasizes the importance of clear, specific, and relevant feedback to improve student learning outcomes.

The usefulness of feedback for students assesses how effective the feedback is in helping students correct mistakes and improve their understanding [46]. Helpful feedback is that which students can immediately apply in subsequent assignments [47]. Effective feedback encourages self-reflection and continuous learning [48].

This research explores the potential for using generative AI as an evaluation and feedback tool in distance legal education. This research is expected to identify the most effective GAI in assessing and providing feedback on law student assignments. The results of this research will provide insight into how GAI can be more thoroughly integrated into the distance learning process to improve the efficiency and quality of assignment assessment and provide constructive feedback for students. Thus, this research has the potential to pave the way for further innovation in the use of AI technology in education, especially law. This research provides practical contributions to educational institutions and enriches academic literature regarding the application of advanced technology in modern learning processes.

2. METHOD

This study aims to evaluate the effectiveness of GAI in providing assessment and feedback on student assignments in the context of distance education in law. The main focus of this study was to assess the accuracy, quality of feedback, and usefulness of feedback provided by five different GAI methods which were then compared with assessments from legal experts. Using a quantitative approach and statistical analysis, this research will provide in-depth insight into how GAI can be optimized in an educational environment.

2.1. Research flow

This study follows a systematic and structured approach, divided into several key stages illustrated in Figure 1. The research process is organized methodically, with each stage clearly defined and sequenced. Refer to Figure 1 to see the primary steps involved in this research.

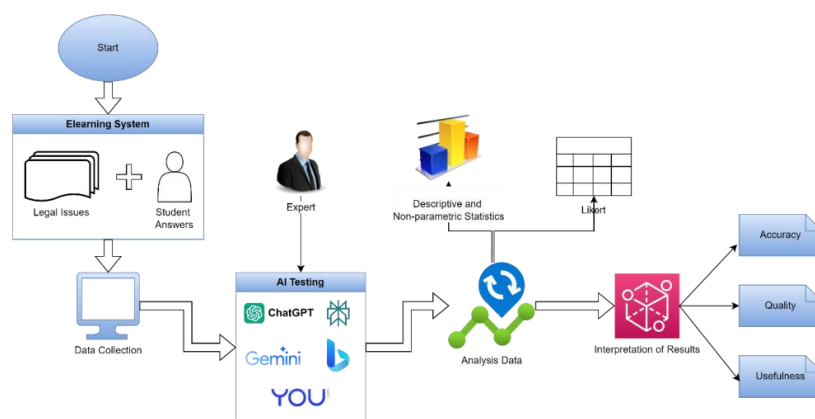


Figure 1. Proposed research flow

This research began with collecting data from the e-learning platform students use to complete their academic assignments. The assignments aim to test students' understanding and skills in law, especially regarding copyright and patent rights. A total of 20 students participated in this research, where they were given questions that required them to analyze hypothetical cases. Students are given sufficient time to work on and collect their answers via the e-learning platform. The selection of 20 students was based on obtaining a sample that was representative enough but could still be managed well in data analysis. Each student submits written answers reflecting their understanding of how technological works can be protected by copyright and patents per relevant regulations and theories.

Once all the answers are collected, the next stage is evaluating the answers using five different GAI. Each GAI assesses student answers based on predetermined criteria, such as accuracy of information, conformity with relevant legal theory, and ability to answer questions comprehensively. The assessment provided by GAI is then compared with the evaluation provided by experienced legal experts. Legal experts assess students' answers using their in-depth knowledge of copyright and patent law, as well as applicable professional standards. This comparison aims to evaluate the extent to which assessments by GAI are in line with expert assessments, as well as to identify significant differences between AI and human evaluations.

Data analysis was carried out using descriptive and non-parametric statistics. Descriptive statistical analysis provides a general description of the data that has been collected, including the calculation of the average and standard deviation of the assessments provided by GAI and experts. For more in-depth analysis,

non-parametric statistical methods are used because this method does not require certain distribution assumptions from the data [49]. The method used includes the Wilcoxon Test to test significant differences between GAI assessments and expert assessments, the intraclass correlation coefficient (ICC) to measure the level of consistency or reliability of assessments between various GAIs and experts, as well as Kappa and Kendall's W to assess the level of agreement between assessments of various GAIs and expert assessments [50], [51]. The accuracy of GAI's assessment is measured by comparing GAI's assessment results with expert assessments using the true metric positive (TP), true negative (TN), false positive (FP), and false negative (FN) [52].

The quality of the feedback provided by each GAI is evaluated by experts using a Likert scale to measure the extent to which the feedback is accurate, relevant, and valuable in an academic context [53]. Students were also asked to rate the usefulness of the feedback they received using a Likert scale, to provide insight into the effectiveness of feedback from various GAIs in the learning context. The questions given to students are designed based on Bloom's Taxonomy, which includes six cognitive levels: remembering, understanding, applying, analyzing, assessing, and creating [54]. This question requires students to identify the basic concepts of copyright and patent rights, explain the difference between copyright and patent rights, use relevant theories and regulations to determine whether work A can be protected by copyright or patent rights, analyze the given case to identify the elements -elements that qualify for copyright and patent protection, assess the validity of the protection that can be afforded to A's work under applicable law, and construct comprehensive and logical arguments supporting their conclusions.

2.2. Testing the quality of feedback according to experts and the usefulness of feedback according to students

To assess the quality of feedback provided by the GAI method, researchers involved five legal experts to conduct an assessment using the Likert scale. This assessment was carried out using a Likert scale with five levels on a scale of 1 (strongly disagree), 2 (disagree), 3 (neutral), 4 (agree), and 5 (strongly agree) [55]. Experts were asked to assess several aspects of the quality of feedback provided by GAI, as stated by [56]. These aspects included clarity (how clear and easy to understand the feedback is), relevance (how relevant the feedback is to the assigned task), depth (how in-depth the analysis and advice provided are), and constructivity (how constructive the feedback is in helping students correct mistakes and improve their understanding).

2.3. Uses of feedback according to students

Students were also asked to evaluate the usefulness of the feedback they received based on several aspects. They assessed the feedback's benefits (how valuable the feedback is in their learning process) and its applicability (how easy the feedback is to apply in future assignments). This evaluation was done using the same five-level Likert scale [57].

2.4. Student assignment questions

The questions given to students to be analyzed and answered by the five GAIs can be seen in Figure 2. After that, the answers were compared with the assessments from experts. In the context of legal education and evaluating student assignments using GAI, the focus lies in the cognitive domain, which includes knowledge and critical thinking skills. This research designed questions based on Bloom's taxonomy to measure various levels of cognitive ability [54], [58].

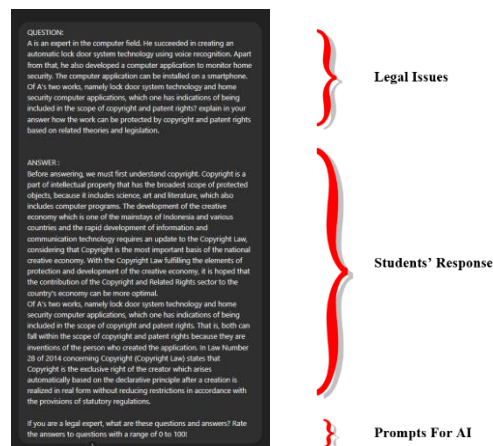


Figure 2. Instructions on AI testing

3. RESULTS

This research aims to explore the effectiveness of GAI as an evaluator and provider of feedback in the context of distance education, with a particular focus on implementing law in legal science. The total respondents in this research were 20 Universitas Terbuka Indonesian students registered in the legal studies study program. Using five types of GAI, this research measures three main aspects: assessment accuracy, feedback quality, and feedback usefulness for students.

3.1. Assessment accuracy analysis

Assessment accuracy is the primary metric used to determine how well GAI meets academic standards in its assessments. This variable is measured in two ways: by comparing GAI assessments with those of legal experts, and by evaluating the true or false results between GAI and legal experts. These approaches help gauge the precision of GAI's evaluations.

3.1.1. Comparison of GAI's assessment with legal experts

In this approach, the assessment given by each GAI is compared with the evaluation given by legal experts. The aim is to see whether GAI can provide judgments that are close to or equivalent to legal experts regarding analysis, conclusions and interpretation of legal cases. Figure 3 shows the assessment results of each GAI compared with assessments from legal experts.

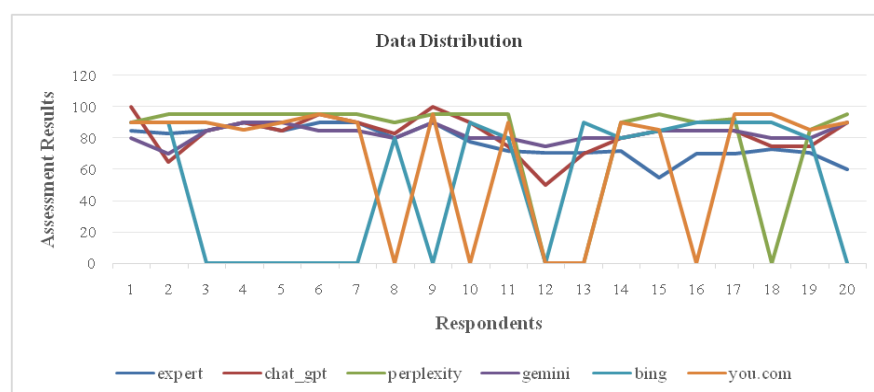


Figure 3. Distribution of GAI assessment data with experts

Based on the data visualization from Figure 3 of the task assessments carried out by various AI methods and compared with the scores given by experts, it can be seen in Figure 3 that GAI tends to give higher assessments than the assessments given by experts. For example, in case number 1, the score given by the expert is 85, while ChatGPT gives a score of 100, Perplexity 90, Gemini 80, Bing 90, and You 90. This shows that the AI method consistently tends to give higher scores. Additionally, there are variations in scoring between different AI methods. For example, on number 4, ChatGPT, Perplexity, and Gemini give a rating of 95, while Bing gives a 90, and You gives an 85. Despite these variations, some GAI methods such as ChatGPT and Gemini show an overall trend level closer to the scores given by experts to each student. The limitations of AI are also visible in cases where the AI method cannot provide a judgment, marked with “maximum character limit reached” or “unable to provide a numerical value”. This suggests that in some situations, AI may encounter difficulties or be unable to provide appropriate assessments, which may affect the reliability of the evaluation process.

a) Descriptive statistical analysis

Descriptive statistical analysis is used to support the analysis results from the data visualization results in Figure 3 to understand how each AI method can provide an assessment compared with the evaluation of experts in legal studies. Descriptive statistical analysis is a method used to describe, show, and summarize data informatively. In the context of this research, descriptive statistical analysis is used to understand the distribution of assessments provided by various artificial intelligence (GAI) models compared to the evaluations of legal experts. This theory involves the use of measures such as the mean (average) and standard deviation to provide a general description of the central tendency and dispersion of assessment data [59]. The mean represents the average of the ratings given, which can help identify how close the AI's assessment is to the expert's assessment. Standard deviation, on the other hand, measures the extent to which the judgments are spread around the mean, which provides insight into the consistency of the judgments.

supplied by each AI method [60]–[62]. This analysis is essential to evaluate whether there is a GAI method that can provide an assessment that is close to expert assessment and to see the consistency of the evaluation provided by each AI method. Table 1 is a presentation of descriptive statistical calculations from the comparison of scores between experts and the five GAI.

Table 1. Descriptive statistical analysis of expert assessments and GAI

Rater	Descriptive statistics	
	Mean	Std. deviation
Expert	77.05	10.15
ChatGPT	82.65	11.96
Perplexity	92.98	2.77
Gemini	82.75	5.25
Bing	86.27	3.67
You	90.33	3.02

Based on the analysis of Table 1 of the average ratings, it can be seen that no artificial intelligence (AI) method reaches or approaches the average rating given by experts, which was recorded at 77.05. In general, all AI methods tend to provide higher ratings compared to expert assessments. Perplexity and You show high-scoring consistency with low standard deviations of 2.77 and 3.02, respectively. This low standard deviation shows that the evaluations from these two methods tend to be stable and consistent in providing results, although the absolute values may not reach the level of assessments given by experts [60], [63]. On the other hand, ChatGPT shows a more significant variation in ratings with a standard deviation of 11.96. This high standard deviation indicates that ratings from ChatGPT have considerable variation, meaning ratings can vary widely depending on the context or question asked. Although this variation may provide flexibility in scoring, it also suggests that consistency in providing grades may be lower than Perplexity and You [64], [65].

To explain this phenomenon, evaluation theory underscores the importance of consistency in assessment to ensure validity and reliability. Low standard deviations, as in Perplexity and You, indicate that although evaluations may not always be perfect according to expert standards, they tend to provide reliable and consistent results [66]. On the other hand, ChatGPT's high standard deviation indicates that although it may provide varying results, there is potential to provide additional insights or broader interpretations of various questions or situations [67]. In evaluating accuracy relative to expert assessments, AI methods that compare average results, such as ChatGPT and Gemini, are recommended because they have average values close to the results provided by experts. The average results are close to the assessment expert's, showing that ChatGPT and Gemini assessments have the same tendency when assessing student assignments.

b) Parametric statistical analysis

In data analysis, non-parametric statistics become relevant when assumptions about data distribution or data characteristics are unmet. Non-parametric statistics does not require data to follow a particular distribution, such as the normal distribution, so it is more flexible to use in various research situations [68], [69]. This method offers a powerful approach to testing hypotheses and measuring relationships between variables without solid assumptions about the shape of the data distribution [70]. The author will explore using the Wilcoxon Test, ICC, and Kappa and Kendall's W in evaluating assessments using GAI in distance education. This analysis will provide deep insight into the consistency, agreement, and differences between AI assessments and expert scores in academic evaluation.

c) Wilcoxon test

In the context of AI evaluation in distance education, the Wilcoxon test is employed to determine if there are significant differences between the assessments given by the five GAIs and those provided by experts. Table 2 presents a Z value and significance (Asymp. Sig. 2-tailed) for each AI-expert pair [71]. The results indicate whether the differences in assessments are statistically significant.

Table 2. Wilcoxon test results

Items	Test statistics ^a				
	ChatGPT-expert	Perplexity-expert	Gemini-expert	Bing-expert	You-expert
Z	-2.070 ^b	-3.928 ^b	-2.203 ^b	-3.103 ^b	-3.696 ^b
Asymp. Sig. (2-tailed)	0.038	0.000	0.028	0.002	0.000

In the calculation results in Table 2, we can conclude that there is a significant difference between the values given by the expert and each AI method (ChatGPT, Perplexity, Gemini, Bing, and You). These

five GAI methods are computational statistics that provide a higher value than the value given by experts. This can be seen from the small p-value (<0.05) and the Z-statistic value which shows the direction of the difference. Thus, these results suggest that AI methods may tend to provide higher assessments than expert assessments.

d) Intraclass correlation coefficient

ICC is a statistical method used to measure the level of consistency or reliability between assessments made by several assessors or measuring tools [72], [73]. In evaluating student assignments, ICC helps determine the extent to which the assessments provided by various AI methods agree with each other and the evaluations provided by experts. In this study, ICC was used to evaluate the reliability of assessments provided by ChatGPT, Perplexity, Gemini, Bing, and You. High ICC values indicate that the assessments of the various GAI methods are consistent with expert assessments, while low ICC values indicate significant variation in the assessments provided. Table 3 presents the ICC results from this study.

Table 3. ICC results

Items	Intraclass correlation ^b	95% Confidence interval	
		Lower bound	Upper bound
Average Measures	.439 ^c	0.094	0.717

ICC value on average measures was 0.439, with a 95% confidence interval [0.094, 0.717], indicating a moderate level of agreement between raters when the average of the ratings was considered. Although there was some consistency, variation in scoring was still significant, indicating the need for improving scoring methods or rater training to achieve higher reliability. The wide range of confidence intervals also indicates uncertainty in these estimates, reinforcing the importance of further refinement.

e) Kappa and Kendall's W

Two non-parametric statistics are used to measure the level of agreement between various GAI methods and expert judgment: Kappa (Cohen's Kappa) and Kendall's W (Kendall's coefficient of concordance). Kappa is a statistical measure that assesses agreement between two or more raters for categorical data, with values ranging from -1 to 1. Negative values indicate more significant disagreement than expected by chance; zero values indicate agreement expected by chance, and positive values indicate higher agreement than expected by chance [74], [75]. Kendall's W, on the other hand, is used to assess agreement between multiple raters for ordinal data, with values ranging from 0 (no agreement) to 1 (perfect agreement) [76]. This analysis is essential to understand the extent to which the various GAI methods align with expert assessments and to assess the consistency of the evaluations provided by the five GAI methods. Table 4 will present the results of Kappa and Kendall's W calculations [77], [78].

Table 4. Kappa and Kendall's W results

Items	Value
Kappa	-0.058
Kendall's W	0.576

The Kappa value of -0.058 shows shallow agreement between expert assessments and the GAI method on student assignments. This indicates that there is a significant difference in the evaluation between experts and GAI. Based on Kendall's W value of 0.576, it shows that there is a pretty good level of agreement between experts and GAI in terms of ranking or preference for student assignments, although not perfect, there is significant consistency in the way they sort or assess student assignments.

3.1.2. Judgment of right and wrong between GAI and experts

A critical aspect of evaluating the accuracy of GAI methods in an educational context is comparing the true or false judgments given by the GAI with the decisions given by experts. This analysis helps understand the extent to which GAI can produce assessments that align with academic standards set by experts. The Table 5 presents data regarding the accuracy of feedback provided by various GAI methods compared with expert judgment for student answers. Each entry in the table indicates whether the assessment provided by each GAI method is correct or incorrect compared to the expert assessment. An accurate evaluation indicates conformity to expert standards, whereas an incorrect appraisal suggests a discrepancy in the evaluation. These data are essential for assessing the ability of various GAI methods to produce accurate and reliable feedback in academic contexts.

Table 5. Crosstab analysis results expert* AI

		ChatGPT			Perplexity			Gemini			Bing			You		
		0	1	Total	0	1	Total	0	1	Total	0	1	Total	0	1	Total
Expert	0	4	7	11	1	10	11	7	4	11	0	11	11	0	11	11
	1	1	8	9	0	9	9	0	9	9	0	9	9	0	9	9
Total		5	15	20	1	19	20	7	13	20	0	20	20	0	20	20

Table 5 shows the results of the crosstab analysis between expert and AI. The test results show that a value of 0 means the answer is wrong, while a value of 1 means the answer is correct. Gemini has higher suitability than the others, where out of 20 answers to student assignments, the expert stated that 11 answers were declared wrong, and 9 answers were declared correct. Gemini stated that 7 answers were stated to be wrong, and 13 answers were asked to be correct. The comparison of the suitability of the assessment of wrong answers between experts and Gemini is 7 student assignments, while the comparison of the suitability of correct answers between experts and Gemini is 9 student assignments.

3.1.3. Accuracy assessment

Assessment accuracy is an important indicator to assess the extent to which the GAI method is reliable in distance education, especially in law. Assessment accuracy shows the percentage of correct assessments produced by the GAI method compared to expert assessments [79]. In this context, high accuracy indicates that the GAI method can provide more consistent assessments and is closer to the standards set by experts. Table 6 will summarize the accuracy of each GAI method evaluated in this study.

Table 6. Accuracy assessment

Items	Accuracy (%)
ChatGPT	60
Perplexity	50
Gemini	80
Bing	45
You	45

Based on Table 6, it can be concluded that Gemini is the method closest to or following the assessment given by experts, with an accuracy rate of 80%. This shows that Gemini can provide more consistent and accurate assessments closer to the standards set by experts. Meanwhile, ChatGPT with an accuracy rate of 60% shows that even though it is not as precise as Gemini, this method is still quite reliable in providing assessments. However, it should be noted that the higher degree of variation in ChatGPT scoring (as indicated by the higher standard deviation in the previous analysis) may have affected its overall reliability. Meanwhile, Perplexity, Bing, and You, with accuracy levels of 50, 45, and 45% respectively, show that this method tends to be less consistent in providing assessments that align with experts. This may be caused by various factors, including the algorithm used and how the method processes and analyzes student assignment data.

3.2. Quality of feedback (expert)

In testing this variable, the quality of the feedback provided by five generative was assessed. The GAI system on student answers was analyzed using a 1-5 Likert scale. Legal experts were asked to evaluate the quality of feedback generated by ChatGPT, Perplexity, Gemini, Bing, and You for 20 student answers. A Likert scale from the statement 1 (strongly disagree) to 5 (strongly agree) is used to measure the extent to which the feedback provided by each GAI is considered accurate, relevant and valuable in the context of legal education [80]. This assessment aims to determine how well each GAI provides quality feedback following academic standards and student learning needs. The data was then analyzed descriptively to obtain an overview of each GAI's average and variability of feedback quality assessments.

Based on the average assessment results given by experts, ChatGPT shows the best performance with an average assessment of 4.22. This indicates that the feedback produced by ChatGPT tends to be more accurate and valuable in the context of legal education. Followed by Gemini which got an average rating of 4.15, showing that GAI also provides good feedback. On the other hand, You got the lowest rating with an average of 3.90. This indicates that You feedback is considered inadequate or not as accurate as other GAIs in assessing student assignments in this context. Bing and Perplexity show intermediate scores with average ratings of 4.09 and 3.99 respectively. Although the feedback from these two GAIs is considered quite good, there is still room for improvement in increasing the accuracy and relevance of their feedback. Thus, these

results illustrate that in the context of evaluating student assignments in the field of legal education, ChatGPT and Gemini can be relied on to provide more accurate feedback compared to You, Bing, and Perplexity.

3.3. Usefulness of feedback (students)

The usefulness of feedback provided by five GAI on student answers was evaluated using a 1-5 Likert scale. Students were asked to rate the extent to which the feedback provided by ChatGPT, Perplexity, Gemini, Bing, and You found them helpful, relevant, and adequate in supporting their learning in the context of legal research. This assessment aims to identify student preferences for the most effective types of feedback and provide insight into how they perceive the quality of the feedback supplied by each GAI. The data will be analyzed to identify the most successful GAI in giving feedback that meets student expectations and needs in online learning.

Based on the test results, students appeared to have varied assessments of the five types of GAI used as feedback tools. ChatGPT received the highest average rating with a score of 4.12, indicating that students tend to feel satisfied with the feedback provided by ChatGPT in the context of their assignments or academic activities. Furthermore, Gemini also received a high rating with an average of 4.07, indicating that students see Gemini as one of the GAIs that is effective in providing relevant and valuable feedback.

On the other hand, Bing and You received lower ratings with an average of 3.91 and 3.80 respectively. This indicates that students may be less satisfied with the feedback provided by both platforms, perhaps due to a lack of depth or relevance of the feedback provided in the context of the material being studied. Despite having a mean of 3.98, Perplexity shows considerable variation in students' ratings, with some students giving low ratings. This shows that although feedback from Perplexity tends to be consistent, some students may feel that the feedback does not always match their expectations or needs in the learning process. Overall, these results indicate the importance of developing and adapting GAI algorithms to provide more consistent and relevant feedback according to students' needs and preferences in distance learning. Further evaluation is also needed to understand more deeply the factors that influence student preferences and satisfaction with various types of feedback provided by GAI.

4. DISCUSSION

GAI has now expanded to various sectors, including higher education [81]. GAI not only helps in content creation and automation of administrative tasks, but also has excellent potential as an evaluator and feedback provider in distance education, especially in legal studies [82], [83]. The potential of AI in education is enormous, from personalizing learning to automating assignment assessments [84], [85]. In the field of education, AI has brought fundamental changes by introducing adaptive learning methods, which can be adapted to the needs and abilities of each student [86], [87]. AI systems can analyze student performance in real time and provide additional material or new challenges according to their needs [24]. Besides that, AI is also used to develop e-learning platforms that enable broader access to education. Specifically in the field of legal science, AI can assist in analyzing legal cases, guide legal research, and even in writing complex legal documents [88], [89].

Previous research has shown that the use of AI in evaluating student assignments in the legal field has great potential. Studies conducted by [90], [91] indicates that some AI methods tend to provide higher assessments than assessments supplied by experts. However, variations in assessment consistency are one of the main challenges faced [92]. This research implies that although AI can provide fast and efficient feedback, there is still a need to improve the consistency and accuracy of the assessments provided. The gap found in previous research is the lack of comprehensive data on how each AI method performs in assessing student assignments in the legal field. This research aims to fill this gap by testing and comparing the accuracy, consistency and relevance of feedback provided by each GAI method using three measurement variable approaches, including accuracy, quality of feedback and usefulness of feedback for students.

4.1. Accuracy of assessment with experts

Accuracy is one of the most crucial factors in evaluation using GAI. Accuracy in this context refers to how much GAI can provide assessments that comply with applicable academic and legal standards. According to studies [39], [40], the accuracy of AI assessments is highly dependent on the algorithm used and the quality of the data used to train the AI. Stated that the accuracy of AI in the evaluation of academic assignments can vary greatly depending on how the data is collected and processed [93]. Research by Liang *et al.* [42] it also supports these findings, showing that AI has great potential to provide highly accurate evaluations if trained with relevant, high-quality data.

In this research, the assessment accuracy variable is categorized into two measurement approaches: comparison of assessments between GAI and legal experts and evaluation of true or false results between

GAI and legal experts. The research results show that the GAI method tends to provide higher assessments than expert assessments. Gemini and ChatGPT show relatively good levels of accuracy with ratings tending to be close to expert scores in some answers. However, variations in scoring were also visible, especially in answers 12 and 13, where there was a significant difference between the scores given by the experts and the scores given by Perplexity (95 vs. 71). These differences highlight the importance of ensuring that AI training data is relevant and covers a wide range of evaluation scenarios to improve assessment accuracy. ChatGPT and Gemini's feedback quality is rated higher than that of other methods, with more detailed and relevant assessments. However, there are also situations where GAI cannot provide an evaluation, as indicated by the "N/A" and "N/B" indicators. This happened to some answers, especially to Bing and You, suggesting that there are limitations in the AI algorithm that may not handle all types of questions or contexts well.

Based on the results of the Wilcoxon test, there is a significant difference between the assessments given by the five GAI methods (ChatGPT, Perplexity, Gemini, Bing, and You) and the assessments offered by experts. The GAI method tends to provide statistically significantly higher ratings than expert assessments, indicating GAI's potential to provide a more positive evaluation of student performance. The ICC analysis shows a moderate level of consistency between the assessments of the various GAI methods when the average of the assessments is considered. Despite this, variation in scoring was still significant, indicating the need for improving scoring methods or rater training to achieve higher reliability. Assessment using Kappa and Kendall's W shows low agreement between expert assessments and the GAI method, especially indicated by the negative value of Kappa. However, Kendall's W value shows consistency in how various GAI methods sort or grade student assignments, even though it is not perfect.

The Crosstab table shows that Gemini gave the same accurate and false ratings as experts with 16 answers, ChatGPT had 12 answers, Perplexity had 10 answers, and Bing and You had 9 answers. This shows that Gemini's and ChatGPT's true and false judgments more comprehensively approach experts' true and false judgments. Although none of the GAI methods achieved expert judgment accuracy, Gemini showed the closest accuracy rate at 80%. ChatGPT, while not as accurate as Gemini, is still quite reliable with a 60% accuracy rate. However, higher variations in ChatGPT scoring are worth noting, as they may affect its overall reliability. These results indicate that although GAI has excellent potential in the academic evaluation process, several limitations must be considered. The reliability and consistency of AI in providing assessments still require improvement, mainly to ensure that the evaluations comply with the academic standards applied by experts. Additionally, AI's limitations in giving judgment in certain situations highlight the need for further development of the algorithms and training data used.

4.2. Feedback quality

Apart from accuracy, GAI's feedback quality is also essential in educational evaluation. Feedback quality involves how insightful and helpful the feedback is for students. High-quality feedback points out errors and provides explanations that help students understand the correct concepts. Emphasize the importance of clear, specific, relevant feedback to improve student learning outcomes [43], [44]. In this research, feedback from GAI is assessed based on how much feedback can help students understand the material better and correct their mistakes.

Based on the average assessment results given by experts, ChatGPT shows the best performance with an average assessment of 4.22. This indicates that the feedback generated by ChatGPT is considered the most accurate and valuable in legal education. For example, in answer number 1, ChatGPT received a score of 4.8 from experts, indicating that the feedback provided by ChatGPT is considered very insightful and relevant. ChatGPT's consistency in delivering high-quality feedback across answers shows its potential as an effective tool in helping students understand and correct their mistakes. Gemini follows with an average rating of 4.15. These results show that the feedback provided by Gemini is also considered quite good by experts. On some answers, such as numbers 2 and 5, Gemini received high marks, 4.6 each, which shows its ability to provide accurate and valuable feedback. With almost equivalent performance to ChatGPT, Gemini can be a reliable alternative for giving feedback on student assignments in the legal field. Bing and Perplexity showed intermediate scores with average ratings of 4.09 and 3.99, respectively. Although the feedback from both GAIs was considered quite good, there were several answers to the question of where the quality of their feedback could be improved. For example, in answer number 1, Bing scored 3.8, which shows that its feedback is still less in-depth than ChatGPT or Gemini. Perplexity, with the second lowest average value, also showed variability in the quality of its feedback, indicating the need for improvements in its algorithm or training data. You received the lowest rating with an average of 3.90, indicating that the feedback provided by You was considered inadequate or not as accurate as other GAIs in assessing student assignments. On some answers, such as numbers 12 and 20, You scored as low as 3.2, indicating that the feedback was often not insightful or relevant enough to help students correct their mistakes. This highlights You's limitations in the context of academic evaluation and the need for significant improvement to compete with other GAIs.

4.3. Usefulness of feedback

The usefulness of feedback for students is the third aspect measured in this research. The usefulness of feedback refers to how effective it is in helping students correct mistakes and improve their understanding. Morris *et al.* [46] stated that valuable feedback is what students can immediately apply to their next assignment. According to Stevens and Levi [47], it supports this view, stating that effective feedback should encourage self-reflection and continuous learning. Yan and Carless [48] added that effective feedback must motivate students to continue learning and improving themselves. In this study, the quality of feedback provided by five GAI is evaluated from the student's perspective. ChatGPT stands out with the highest rating average of 4.12, indicating high student satisfaction with the feedback provided. Gemini also received good ratings with an average of 4.07, suggesting that students see it as an effective tool for providing relevant feedback. On the other hand, Bing and You received lower ratings with averages of 3.91 and 3.80, indicating a lack of depth or relevance in their feedback. Despite having a respectable average of 3.98, Perplexity showed variation in student satisfaction, with some scoring lower than expected.

5. CONCLUSION

This research demonstrates the potential of GAI as an evaluator and feedback provider in distance legal education, emphasizing accuracy, feedback quality, and usability as key variables. Accuracy in binary judgements varied, with Gemini reaching 80%, ChatGPT 60%, and other GAIs having lower scores (Perplexity 50%, Bing 45%, You 45%). In addition, there were significant differences between expert and GAI scores, with ChatGPT often giving higher scores, as seen in case number 1, where the expert gave a score of 85, while ChatGPT gave 100, Perplexity 90, Gemini 80, Bing 90, and You 90. Further statistical analysis revealed moderate agreement (ICC=0.439) and low alignment (Kappa=-0.058) between GAI and expert evaluations, while a Kendall's W value of 0.576 indicated moderate consistency in ratings. On the ratings of feedback quality and feedback usability, ChatGPT emerged as the most effective GAI, expert (4.22) and student (4.12), followed by Gemini with ratings of 4.15 and 4.07. Thus, the selection of appropriate GAIs such as ChatGPT and Gemini has the potential to improve the quality of student learning in the context of evaluation of academic assignments and activities. This research emphasizes the importance of considering the characteristics and performance of GAIs in supporting online learning processes, particularly in legal education, to maximize the benefits of technology in education. This research supports previous theories that AI technology can improve the quality of education by providing more objective assessment and more useful feedback. The implications of this research suggest that by continuing to develop and refine GAI technology, this tool can be very useful in improving the quality of distance legal education. Further research is needed to explore the full potential of GAI and overcome existing challenges.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to Universitas Terbuka for the support provided in the execution of this research. Although this study was funded through independent research funding, the assistance from Universitas Terbuka through the research and community service institute (LPPM-UT) has been essential to the successful completion of this project. We are deeply appreciate Universitas Terbuka's commitment to fostering academic research.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dian Nurdiana	✓	✓	✓	✓	✓	✓			✓	✓		✓	✓	✓
Muhamad Riyan Maulana	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Siti Hadijah Hasanah		✓		✓	✓	✓	✓	✓	✓	✓				
Madiha Dzakiyyah Chairunnisa						✓	✓	✓		✓	✓			
Avelyn Pingkan Komuna						✓	✓	✓		✓	✓			
Muhammad Rif'an						✓				✓	✓			

C : Conceptualization
M : Methodology
So : Software
Va : Validation
Fo : Formal analysis

I : Investigation
R : Resources
D : Data Curation
O : Writing - Original Draft
E : Writing - Review & Editing

Vi : Visualization
Su : Supervision
P : Project administration
Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

The research related to human use has been complied with all the relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board or equivalent committee.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [DN], upon reasonable request.

REFERENCES

- [1] H. Al Naqbi, Z. Bahroun, and V. Ahmed, "Enhancing work productivity through generative artificial intelligence: a comprehensive literature review," *Sustainability*, vol. 16, no. 3, Jan. 2024, doi: 10.3390/su16031166.
- [2] A. K. Kar, P. S. Varsha, and S. Rajan, "Unravelling the impact of generative artificial intelligence (GAI) in industrial applications: a review of scientific and grey literature," *Global Journal of Flexible Systems Management*, vol. 24, no. 4, pp. 659–689, Dec. 2023, doi: 10.1007/s40171-023-00356-x.
- [3] Y.-C. Hsu and Y.-H. Ching, "Generative artificial intelligence in education, part one: the dynamic frontier," *TechTrends*, vol. 67, no. 4, pp. 603–607, Jul. 2023, doi: 10.1007/s11528-023-00863-9.
- [4] K.-B. Ooi *et al.*, "The potential of generative artificial intelligence across disciplines: perspectives and future directions," *Journal of Computer Information Systems*, Oct. 2023, doi: 10.1080/08874417.2023.2261010.
- [5] W. Lo, C.-M. Yang, Q. Zhang, and M. Li, "Increased productivity and reduced waste with robotic process automation and generative AI-powered IOE services," *Journal of Web Engineering*, vol. 23, no. 1, pp. 53–88, Mar. 2024, doi: 10.13052/jwe1540-9589.2313.
- [6] P. Lameris and S. Arnab, "Power to the teachers: an exploratory review on artificial intelligence in education," *Information*, vol. 13, no. 1, Dec. 2021, doi: 10.3390/info13010014.
- [7] Z. Bahroun, C. Anane, V. Ahmed, and A. Zacca, "Transforming education: a comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis," *Sustainability*, vol. 15, no. 17, Aug. 2023, doi: 10.3390/su151712983.
- [8] E. Đerić, D. Frank, and M. Malenica, "Comparison and quantification of GAI tools use among different academic population segments," in *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, 2024, pp. 730–735, doi: 10.1109/MIPRO60963.2024.10569253.
- [9] K. Sevnarayan and M.-A. Potter, "Generative artificial intelligence in distance education: transformations, challenges, and impact on academic integrity and student voice," *Journal of Applied Learning & Teaching*, vol. 7, no. 1, Apr. 2024, doi: 10.37074/jalt.2024.7.1.41.
- [10] K. Rybinski and E. Kopciuszewska, "Will artificial intelligence revolutionise the student evaluation of teaching? a big data study of 1.6 million student reviews," *Assessment & Evaluation in Higher Education*, vol. 46, no. 7, pp. 1127–1139, Oct. 2021, doi: 10.1080/02602938.2020.1844866.
- [11] M. Liu, Y. Ren, L. M. Nyagoga, F. Stonier, Z. Wu, and L. Yu, "Future of education in the era of generative artificial intelligence: consensus among chinese scholars on applications of ChatGPT in schools," *Future in Educational Research*, vol. 1, no. 1, pp. 72–101, Sep. 2023, doi: 10.1002/fer3.10.
- [12] I. Pesovski, R. Santos, R. Henriques, and V. Trajkovic, "Generative AI for customizable learning experiences," *Sustainability*, vol. 16, no. 7, Apr. 2024, doi: 10.3390/su16073034.
- [13] D. Griffiths, E. Frías-Martínez, A. Tlili, and D. Burgos, "A cybernetic perspective on generative AI in education: from transmission to coordination," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 5, 2024, doi: 10.9781/ijimai.2024.02.008.
- [14] K. Nikolopoulou, "Generative artificial intelligence in higher education: exploring ways of harnessing pedagogical practices with the assistance of ChatGPT," *International Journal of Changes in Education*, vol. 1, no. 2, pp. 103–111, Apr. 2024, doi: 10.47852/bonviewIJCE42022489.
- [15] S. Balat, M. Yavuz, and B. Kayalı, "Generative artificial intelligence as academic assistant," in *Transforming Education With Generative AI: Prompt Engineering and Synthetic Content Creation*, IGI Global, 2024, pp. 138–157, doi: 10.4018/979-8-3693-1351-0.ch007.
- [16] S. J. Shi, J. W. Li, and R. Zhang, "A study on the impact of generative artificial intelligence supported situational interactive teaching on students' 'flow' experience and learning effectiveness — a case study of legal education in china," *Asia Pacific Journal of Education*, vol. 44, no. 1, pp. 112–138, Jan. 2024, doi: 10.1080/02188791.2024.2305161.

Generative artificial intelligence as an evaluator and feedback tool in distance learning: ... (Dian Nurdiana)




- [17] W. Alschnner, "The computational analysis of international law," in *Research Methods in International Law*, vol. 3, no. 1, Edward Elgar Publishing, 2021, pp. 8–17, doi: 10.4337/9781788972369.00022.
- [18] P. Grajzl and P. Murrell, "Caselaw and England's economic performance during the industrial revolution: data and evidence," *Journal of Comparative Economics*, vol. 52, no. 1, pp. 145–165, Mar. 2024, doi: 10.1016/j.jce.2023.10.002.
- [19] A. Shevchenko, S. Kydin, S. Kamarali, and M. Dei, "Issues with interpreting the social and legal value of a person in the context of the integrative type of legal-awareness," *Fundamental and applied researches in practice of leading scientific schools*, vol. 38, no. 2, pp. 54–61, Apr. 2020, doi: 10.33531/farplss.2020.2.10.
- [20] R. A. Cass, C. S. Diver, J. M. Beermann, and J. L. Mascott, *Administrative law: cases and materials*. Aspen Publishing, 2024.
- [21] R. Wacks, *Understanding jurisprudence: an introduction to legal theory*. Oxford University Press, 2020.
- [22] J. C. Dernbach, R. V. Singleton, C. S. Wharton, C. J. Wasson, and J. M. Ruhtenberg, *A practical guide to legal writing and legal method*. Aspen Publishing, 2021.
- [23] L. C. Oates, A. Enquist, and J. Francis, *The legal writing handbook: analysis, research, and writing*. Aspen Publishing, 2021.
- [24] M. Hooda, C. Rana, O. Dahiya, A. Rizwan, and M. S. Hossain, "Artificial intelligence for assessment and feedback to enhance student success in higher education," *Mathematical Problems in Engineering*, vol. 2022, no. 1, May 2022, doi: 10.1155/2022/5215722.
- [25] T. Nazaretsky, M. Cukurova, and G. Alexandron, "An instrument for measuring teachers' trust in AI-based educational technology," in *LAK22: 12th International Learning Analytics and Knowledge Conference*, New York, USA: ACM, Mar. 2022, pp. 56–66, doi: 10.1145/3506860.3506866.
- [26] M. Messer, N. C. C. Brown, M. Kölling, and M. Shi, "Automated grading and feedback tools for programming education: a systematic review," *ACM Transactions on Computing Education*, vol. 24, no. 1, pp. 1–43, Mar. 2024, doi: 10.1145/3636515.
- [27] J. C. Sánchez-Prieto, J. Cruz-Benito, R. Therón, and F. García-Peñalvo, "Assessed by machines: development of a TAM-based tool to measure AI-based assessment acceptance among students," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 4, 2020, doi: 10.9781/ijimai.2020.11.009.
- [28] A. Han et al., "Teachers, parents, and students' perspectives on integrating generative AI into elementary literacy education," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, New York, USA: ACM, May 2024, pp. 1–17, doi: 10.1145/3613904.3642438.
- [29] R. M. -Villarreal, E. V. Perdomo, D. E. S. -Navarro, R. T. -Aguilera, and F. S. Gerardou, "Challenges and opportunities of generative AI for higher education as explained by chatGPT," *Education Sciences*, vol. 13, no. 9, Aug. 2023, doi: 10.3390/educsci13090856.
- [30] S. Mallik and A. Gangopadhyay, "Proactive and reactive engagement of artificial intelligence methods for education: a review," *Frontiers in Artificial Intelligence*, vol. 6, May 2023, doi: 10.3389/frai.2023.1151391.
- [31] J. Mao, B. Chen, and J. C. Liu, "Generative artificial intelligence in education and its implications for assessment," *TechTrends*, vol. 68, no. 1, pp. 58–66, Jan. 2024, doi: 10.1007/s11528-023-00911-4.
- [32] H. Yu and Y. Guo, "Generative artificial intelligence empowers educational reform: current status, issues, and prospects," *Frontiers in Education*, vol. 8, Jun. 2023, doi: 10.3389/feduc.2023.1183162.
- [33] C. Zastudil, M. Rogalska, C. Kapp, J. Vaughn, and S. MacNeil, "Generative AI in computing education: perspectives of students and instructors," in *2023 IEEE Frontiers in Education Conference (FIE)*, 2023, pp. 1–9, doi: 10.1109/FIE58773.2023.10343467.
- [34] B. Anthony Jnr and S. Noel, "Examining the adoption of emergency remote teaching and virtual learning during and after COVID-19 pandemic," *International Journal of Educational Management*, vol. 35, no. 6, pp. 1136–1150, Oct. 2021, doi: 10.1108/IJEM-08-2020-0370.
- [35] G. Öztüdoğlu, "Problems faced in distance education during COVID-19 pandemic," *Participatory Educational Research*, vol. 8, no. 4, pp. 321–333, Dec. 2021, doi: 10.17275/per.21.92.8.4.
- [36] A. Y. Q. Huang, O. H. T. Lu, and S. J. H. Yang, "Effects of artificial intelligence-enabled personalized recommendations on learners' learning engagement, motivation, and outcomes in a flipped classroom," *Computers & Education*, vol. 194, Mar. 2023, doi: 10.1016/j.compedu.2022.104684.
- [37] B. L. Moorhouse, M. A. Yeo, and Y. Wan, "Generative AI tools and assessment: guidelines of the world's top-ranking universities," *Computers and Education Open*, vol. 5, Dec. 2023, doi: 10.1016/j.caeo.2023.100151.
- [38] J. Su and W. Yang, "Unlocking the power of ChatGPT: a framework for applying generative ai in education," *ECNU Review of Education*, vol. 6, no. 3, pp. 355–366, Aug. 2023, doi: 10.1177/20965311231168423.
- [39] K. Moulai, A. Yadegari, M. Baharestani, S. Farzanbakhsh, B. Sabet, and M. R. Afrash, "Generative artificial intelligence in healthcare: a scoping review on benefits, challenges and applications," *International Journal of Medical Informatics*, vol. 188, Aug. 2024, doi: 10.1016/j.ijmedinf.2024.105474.
- [40] R. Raman, P. Calyam, and K. Achuthan, "ChatGPT or bard: who is a better certified ethical hacker?," *Computers & Security*, vol. 140, 2024, doi: 10.1016/j.cose.2024.103804.
- [41] T. Schaffter et al., "Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms," *JAMA Network Open*, vol. 3, no. 3, Mar. 2020, doi: 10.1001/jamanetworkopen.2020.0265.
- [42] W. Liang et al., "Advances, challenges and opportunities in creating data for trustworthy AI," *Nature Machine Intelligence*, vol. 4, no. 8, pp. 669–677, Aug. 2022, doi: 10.1038/s42256-022-00516-1.
- [43] E. Panadero and A. A. Lipnevich, "A review of feedback models and typologies: towards an integrative model of feedback elements," *Educational Research Review*, vol. 35, Feb. 2022, doi: 10.1016/j.edurev.2021.100416.
- [44] S. Tong, N. Jia, X. Luo, and Z. Fang, "The janus face of artificial intelligence feedback: deployment versus disclosure effects on employee performance," *Strategic Management Journal*, vol. 42, no. 9, pp. 1600–1631, Sep. 2021, doi: 10.1002/smj.3322.
- [45] A. A. Lipnevich and E. Panadero, "A review of feedback models and theories: descriptions, definitions, and conclusions," *Frontiers in Education*, vol. 6, Dec. 2021, doi: 10.3389/feduc.2021.720195.
- [46] R. Morris, T. Perry, and L. Wardle, "Formative assessment and feedback for learning in higher education: a systematic review," *Review of Education*, vol. 9, no. 3, 2021, doi: 10.1002/rev3.3292.
- [47] D. D. Stevens and A. J. Levi, *Introduction to rubrics*. New York: Routledge, 2023, doi: 10.4324/9781003445432.
- [48] Z. Yan and D. Carless, "Self-assessment is about more than self: the enabling role of feedback literacy," *Assessment & Evaluation in Higher Education*, vol. 47, no. 7, pp. 1116–1128, Oct. 2022, doi: 10.1080/02602938.2021.2001431.
- [49] E. B. Manoukian, *Mathematical nonparametric statistics*. Taylor & francis, 2022.
- [50] I. Navarro-Soria, J. R. Rico-Juan, R. Juárez-Ruiz de Mier, and R. Lavigne-Cervan, "Prediction of attention deficit hyperactivity disorder based on explainable artificial intelligence," *Applied Neuropsychology: Child*, pp. 1–14, Apr. 2024, doi: 10.1080/21622965.2024.2336019.
- [51] R.-M. Pan, H.-J. Chang, M.-J. Chi, C.-Y. Wang, and Y.-H. Chuang, "The traditional chinese version of the geriatric anxiety inventory: psychometric properties and cutoff point for detecting anxiety," *Geriatric Nursing*, vol. 58, pp. 438–445, Jul. 2024, doi: 10.1016/j.gerinurse.2024.06.008.

- [52] S. Parasa, A. Repici, T. Berzin, C. Leggett, S. A. Gross, and P. Sharma, "Framework and metrics for the clinical use and implementation of artificial intelligence algorithms into endoscopy practice: recommendations from the american society for gastrointestinal endoscopy artificial intelligence task force," *Gastrointestinal Endoscopy*, vol. 97, no. 5, pp. 815-824, May 2023, doi: 10.1016/j.gie.2022.10.016.
- [53] K.-D. Vattoy, S. M. Gamlem, and W. M. Rogne, "Examining students' feedback engagement and assessment experiences: a mixed study," *Studies in Higher Education*, vol. 46, no. 11, pp. 2325-2337, Nov. 2021, doi: 10.1080/03075079.2020.1723523.
- [54] M. D. Laddha, V. T. Lokare, A. W. Kiwelekar, and L. D. Netak, "Classifications of the summative assessment for revised bloom's taxonomy by using deep learning," *International Journal of Engineering Trends and Technology*, vol. 69, no. 3, pp. 211-218, Mar. 2021, doi: 10.14445/22315381/IJETT-V69I3P232.
- [55] A. T. Alabi and M. O. Jelili, "Clarifying Likert scale misconceptions for improved application in urban studies," *Quality & Quantity*, vol. 57, no. 2, pp. 1337-1350, Apr. 2023, doi: 10.1007/s11135-022-01415-8.
- [56] M. Pande and S. V. Bharathi, "Theoretical foundations of design thinking – a constructivism learning approach to design thinking," *Thinking Skills and Creativity*, vol. 36, Jun. 2020, doi: 10.1016/j.tsc.2020.100637.
- [57] Y. Liu, W. Xiong, Y. Xiong, and Y. B. Wu, "Generating timely individualized feedback to support student learning of conceptual knowledge in writing-to-learn activities," *Journal of Computers in Education*, vol. 11, no. 2, pp. 367-399, Jun. 2024, doi: 10.1007/s40692-023-00261-3.
- [58] A. Momen, M. Ebrahimi, and A. M. Hassan, "Importance and implications of theory of bloom's taxonomy in different fields of education," in *International conference on emerging technologies and intelligent systems*, Springer, 2023, pp. 515-525, doi: 10.1007/978-3-031-20429-6_47.
- [59] C. A. Mertler, R. A. Vannatta, and K. N. LaVenja, *Advanced and multivariate statistical methods*. New York: Routledge, 2021, doi: 10.4324/9781003047223.
- [60] S. McGrath *et al.*, "Estimating the sample mean and standard deviation from commonly reported quantiles in meta-analysis," *Statistical Methods in Medical Research*, vol. 29, no. 9, pp. 2520-2537, Sep. 2020, doi: 10.1177/0962280219889080.
- [61] H. N. Amer, N. Y. Dahlan, A. M. Azmi, M. F. A. Latip, M. S. Onn, and A. Tumian, "Solar power prediction based on artificial neural network guided by feature selection for large-scale solar photovoltaic plant," *Energy Reports*, vol. 9, pp. 262-266, Nov. 2023, doi: 10.1016/j.egy.2023.09.141.
- [62] J. Shi *et al.*, "Optimally estimating the sample standard deviation from the five-number summary," *Research Synthesis Methods*, vol. 11, no. 5, pp. 641-654, Sep. 2020, doi: 10.1002/jrsm.1429.
- [63] W.-H. Huang, "Control charts for joint monitoring of the lognormal mean and standard deviation," *Symmetry*, vol. 13, no. 4, Mar. 2021, doi: 10.3390/sym13040549.
- [64] B. Stoler and A. Nekrutenko, "Sequencing error profiles of illumina sequencing instruments," *NAR Genomics and Bioinformatics*, vol. 3, no. 1, Jan. 2021, doi: 10.1093/nargab/lqab019.
- [65] L. da Conceição Braga, B. Ô. P. Gonçalves, P. L. Coelho, A. L. da Silva Filho, and L. M. Silva, "Identification of best housekeeping genes for the normalization of RT-QPCR in human cell lines," *Acta Histochemica*, vol. 124, no. 1, Jan. 2022, doi: 10.1016/j.acthis.2021.151821.
- [66] Y. Dzakadzie and F. Quansah, "Modeling unit non-response and validity of online teaching evaluation in higher education using generalizability theory approach," *Frontiers in Psychology*, vol. 14, Sep. 2023, doi: 10.3389/fpsyg.2023.1202896.
- [67] L. Lohman, "Evaluation of university teaching as sound performance appraisal," *Studies in Educational Evaluation*, vol. 70, Sep. 2021, doi: 10.1016/j.stueduc.2021.101008.
- [68] E. S. Mukasa, W. Christospher, B. Ivan, and M. Kizito, "The effects of parametric, non-parametric tests and processes in inferential statistics for business decision making," *Open Journal of Business and Management*, vol. 9, no. 3, pp. 1510-1526, 2021, doi: 10.4236/ojbm.2021.93081.
- [69] A. Moradi-Motlagh and A. Emrouznejad, "The origins and development of statistical approaches in non-parametric frontier models: a survey of the first two decades of scholarly literature (1998-2020)," *Annals of Operations Research*, vol. 318, no. 1, pp. 713-741, Nov. 2022, doi: 10.1007/s10479-022-04659-7.
- [70] P. H. Westfall and A. L. Arias, *Understanding regression analysis*. Boca Raton : CRC Press: Chapman and Hall/CRC, 2020, doi: 10.1201/9781003025764.
- [71] D. Berrar, "Using p-values for the comparison of classifiers: pitfalls and alternatives," *Data Mining and Knowledge Discovery*, vol. 36, no. 3, pp. 1102-1139, May 2022, doi: 10.1007/s10618-022-00828-1.
- [72] Y. Takahashi, Y. Fujino, K. Miura, A. Toida, T. Matsuda, and S. Makita, "Intra- and inter-rater reliability of rectus femoris muscle thickness measured using ultrasonography in healthy individuals," *The Ultrasound Journal*, vol. 13, no. 1, Dec. 2021, doi: 10.1186/s13089-021-00224-8.
- [73] M. Groenier, L. Brummer, B. P. Bunting, and A. G. Gallagher, "Reliability of observational assessment methods for outcome-based assessment of surgical skill: systematic review and meta-analyses," *Journal of Surgical Education*, vol. 77, no. 1, pp. 189-201, Jan. 2020, doi: 10.1016/j.jsurg.2019.07.007.
- [74] D. Chicco, M. J. Warrens, and G. Jurman, "The matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment," *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3084050.
- [75] M. J. Warrens, "Kappa coefficients for dichotomous-nominal classifications," *Advances in Data Analysis and Classification*, vol. 15, no. 1, pp. 193-208, Mar. 2021, doi: 10.1007/s11634-020-00394-8.
- [76] A. de Raadt, M. J. Warrens, R. J. Bosker, and H. A. L. Kiers, "A comparison of reliability coefficients for ordinal rating scales," *Journal of Classification*, vol. 38, no. 3, pp. 519-543, Oct. 2021, doi: 10.1007/s00357-021-09386-5.
- [77] A. S. Kolesnyk and N. F. Khairova, "Justification for the use of Cohen's Kappa statistic in experimental studies of NLP and text mining," *Cybernetics and Systems Analysis*, vol. 58, no. 2, pp. 280-288, Mar. 2022, doi: 10.1007/s10559-022-00460-3.
- [78] M. Nasser *et al.*, "Feature reduction for molecular similarity searching based on autoencoder deep learning," *Biomolecules*, vol. 12, no. 4, Mar. 2022, doi: 10.3390/biom12040508.
- [79] E. Yüzer, V. Doğan, V. Kılıç, and M. Şen, "Smartphone embedded deep learning approach for highly accurate and automated colorimetric lactate analysis in SWEAT," *Sensors and Actuators B: Chemical*, vol. 371, Nov. 2022, doi: 10.1016/j.snb.2022.132489.
- [80] C. León-Mantero, J. C. Casas-Rosal, C. Pedrosa-Jesús, and A. Maz-Machado, "Measuring attitude towards mathematics using Likert scale surveys: the weighted average," *PLoS ONE*, vol. 15, pp. 1-15, 2020, doi: 10.1371/journal.pone.0239626.
- [81] T. Farrelly and N. Baker, "Generative artificial intelligence: implications and considerations for higher education practice," *Education Sciences*, vol. 13, no. 11, Nov. 2023, doi: 10.3390/educsci13111109.
- [82] T. R. Sales de Aguiar, "ChatGPT: reflections from the UK higher education institutions, accountancy bodies and big4s," *Accounting Research Journal*, vol. 37, no. 3, pp. 308-329, Jul. 2024, doi: 10.1108/ARJ-07-2023-0184.




- [83] R. Watermeyer, L. Phipps, D. Lanclos, and C. Knight, "Generative ai and the automating of academia," *Postdigital Science and Education*, vol. 6, no. 2, pp. 446–466, Jun. 2024, doi: 10.1007/s42438-023-00440-6.
- [84] V. González-Calatayud, P. Prendes-Espinosa, and R. Roig-Vila, "Artificial intelligence for student assessment: a systematic review," *Applied Sciences*, vol. 11, no. 12, Jun. 2021, doi: 10.3390/app11125467.
- [85] K. K. Wong, "Blended learning and AI: enhancing teaching and learning in higher education," in *International Conference on Blended Learning*, Springer, 2024, pp. 39–61, doi: 10.1007/978-981-97-4442-8_4.
- [86] T. Kabudi, I. Pappas, and D. H. Olsen, "AI-enabled adaptive learning systems: a systematic mapping of the literature," *Computers and Education: Artificial Intelligence*, vol. 2, 2021, doi: 10.1016/j.caeai.2021.100017.
- [87] I. Gligorea, M. Cioca, R. Oancea, A.-T. Gorski, H. Gorski, and P. Tudorache, "Adaptive learning using artificial intelligence in e-learning: a literature review," *Education Sciences*, vol. 13, no. 12, Dec. 2023, doi: 10.3390/educsci13121216.
- [88] K. Atkinson, T. Bench-Capon, and D. Bollegala, "Explanation in AI and law: past, present and future," *Artificial Intelligence*, vol. 289, 2020, doi: 10.1016/j.artint.2020.103387.
- [89] K. Ahuja and I. Bala, "Role of artificial intelligence and iot in next generation education system," in *Intelligence of Things: AI-IoT Based Critical-Applications and Innovations*, Cham: Springer International Publishing, 2021, pp. 189–208, doi: 10.1007/978-3-030-82800-4_8.
- [90] E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: review of empirical research," *Academy of Management Annals*, vol. 14, no. 2, pp. 627–660, Jul. 2020, doi: 10.5465/annals.2018.0057.
- [91] J. Wang, M. D. Molina, and S. S. Sundar, "When expert recommendation contradicts peer opinion: relative social influence of valence, group identity and artificial intelligence," *Computers in Human Behavior*, vol. 107, Jun. 2020, doi: 10.1016/j.chb.2020.106278.
- [92] R. Preston, M. Gratani, K. Owens, P. Roche, M. Zimanyi, and B. Malau-Aduli, "Exploring the impact of assessment on medical students' learning," *Assessment & Evaluation in Higher Education*, vol. 45, no. 1, pp. 109–124, Jan. 2020, doi: 10.1080/02602938.2019.1614145.
- [93] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York, NY, USA: ACM, Jan. 2020, pp. 295–305, doi: 10.1145/3351095.3372852.

BIOGRAPHIES OF AUTHORS






Dian Nurdiana    is Assistant Professor and is one of the lecturers in the Information Systems Study Program at Universitas Terbuka. He was appointed as a lecturer at the university in 2019. He was appointed as an Assistant Professor in 2020. Interests are in information systems, software testing, and technology-related learning. He can be contacted at email: dian.nurdiana@ecampus.ut.ac.id.






Muhamad Riyan Maulana    is a final-year student of the Information Systems Study Program at Universitas Terbuka, currently awaiting his graduation ceremony. Throughout his studies, he actively participated in student activities, including competitions, and was deeply involved in research and community service. His dedication earned him the title of Universitas Terbuka's Top Outstanding Student (*Mahasiswa Berprestasi/Mawapres Utama I*) in both 2023 and 2024, all while managing a full-time job. As a student, he authored 21 scientific articles published in SINTA-accredited journals (ranked 2, 3, and 4 by the Ministry of Education and Culture of the Republic of Indonesia) and internationally indexed journals such as Copernicus. His research interests include information systems and software testing, data mining, artificial intelligence, decision support systems, and the implementation of technology in educational learning processes. He can be contacted at email: 042904491@ecampus.ut.ac.id.






Siti Hadijah Hasanah    is a teaching staff at the Faculty of Science and Technology, Statistics Study Program of the Universitas Terbuka. She graduated from the Faculty of Science and Technology of Syarif Hidayatullah State Islamic University in Mathematics in 2008. Then obtained a master's degree of science from the Bogor Agricultural Institute in Statistics in 2015. Her research interests are statistics and computer science in distance education. She can be contacted at email: sitihadijah@ecampus.ut.ac.id.






Madiha Dzakiyyah Chairunnisa    received a Masters in Law from Gadjah Mada University. She has more than 10 years of experience as an academic. She is also registered as a certified mediator. In her daily life, she is a lecturer at the Faculty of Law, Social Sciences and Political Sciences. Her current research interests are in the field of intellectual property rights, arbitration and dispute resolution both national and international, metaverse development in the legal field. Her publication topics include the development of virtual reality in legal education, law and intellectual property rights. She can be contacted at email: madiha.chairunnisa@ecampus.ut.ac.id.



Avelyn Pingkan Komuna    graduated from the Faculty of Law at Hasanuddin University and she is currently a lecturer at the Law Study Program at Universitas Terbuka, Indonesia. She teaches intellectual property law and her research focuses on intellectual property. Additionally, she conducts research in the field of distance learning development in law. Her research outputs include the civil trial practice guide for distance learning, online moot court practice, and the development of virtual reality trials. She can be contacted at email: avelynkomuna@ecampus.ut.ac.id.



Muhammad Rifan    obtained his doctorate in electrical engineering from the University of Indonesia. With over 20 years of experience, he has worked as an academic, practitioner, and expert in the fields of artificial intelligence and education policy. Currently, he serves as Head Lecturer at the Faculty of Science and Technology, at Universitas Terbuka, and also as Director of Information Systems. His current research interests focus on artificial intelligence and its applications in industry. His publication topics include fuzzy logic, neural networks, intelligent control systems, renewable energy, and sustainable development. He can be contacted at email: m.rifan@ecampus.ut.ac.id.