

How Teacher-educators and ChatGPT Evaluate Pre-service Teachers' Writing? : Focused on Newspaper Editorials

Ahn, Sanghee

Jeonbuk National University
Department of Korean Language Education
Lecturer (1st author)

Kim, Kyoungghwa

Cheongju University
Department of Korean Language Education
Assistant Professor (Co-author)

Kim, Youngran

Kangwon National University
Department of Korean Language Education
Professor (Corresponding author)

- * The main idea of this paper was presented at the 5th Southeast Asian Conference on Education in Malaysia(2025, Feb. 24.).

- I. Introduction
- II. Literature Review
- III. Research Method
- IV. Results
- V. Discussion

I. Introduction

In recent years, the use of artificial intelligence('AI' hereinafter) in the field of education has become a major topic of discussion(Shidiq, 2023; Kolade, 2024). Various attempts have been made to explore or develop ways to support teaching & learning and to assess learners' performance through AI. In particular, many researchers and educators have shown great interest in applying conversational AI services such as ChatGPT to education, and a wide range of studies have been conducted in this area(Imran & Almusharraf, 2023; Tajik, 2024). ChatGPT is a conversational AI chatbot based on the Generative Pre-trained Transformer(GPT) developed by OpenAI and was released in November 2022. ChatGPT analyzes user input and generates appropriate responses, enabling natural conversations as if interacting with a human being.

This study examines the educational applicability of ChatGPT within the context of writing evaluation. The aims of this study can be discussed from two perspectives. First, generative AI such as ChatGPT can be a potential solution to the challenges inherent in writing evaluation. In educational settings, writing should be ideally evaluated by multiple raters to ensure validity and reliability. However, within

the Korean educational context, it is often not feasible to have multiple teachers collaboratively assessing student writing. This difficulty arises from structural constraints—each teacher is responsible for too many students and must handle not only teaching and evaluation but also administrative duties. Moreover, writing evaluation is highly susceptible to subjective interpretation, making it difficult to maintain consistency & reliability in rating, while also requiring significant time and effort.

Within these constraints, ChatGPT, as a large language model, has the potential to enhance the objectivity and efficiency of evaluation by rapidly analyzing learners' writing and generating feedback based on consistent application of criteria (Tajik, 2024). Imran & Almusharrarf (2023) has shown that ChatGPT can assist teachers in assessing tasks. This study seeks to explore the potential and limitations of applying generative AI, such as ChatGPT, to writing evaluation as a means to secure greater objectivity, consistency, and validity in the process of writing evaluation.

Second, despite the potential of ChatGPT in writing evaluation, research on this topic remains scarce in Korea. According to Allam, Dempere, Akre, Parakash, Mazher, & Ahamed (2023), the use of AI has not yet been sufficiently investigated or applied across various educational domains, and because ChatGPT is a relatively new technology, related empirical research is still limited. It is essential to first understand how AI can (or cannot) enhance teaching & learning for both teachers and students.

Existing studies on AI-use in writing evaluation have primarily focused on automated essay scoring (Cha, Choi, & Lee, 2024; Choi, Kim, & Kim, 2025; Son, 2023), feedback generation (Choi, 2025; Lee, 2025a; Son, 2025), and forward-looking discussions on AI in writing education (Jang, 2023; Lee, 2025b), whereas empirical studies on the use of generative AI in rating remain rare. In particular, few studies have compared writing evaluation between ChatGPT and teacher-evaluators, although recent studies in writing education have reported that

ChatGPT can be useful for idea generation, drafting, and feedback provision (AlAfnan, Dishari, Jovic, & Lomidze, 2023; Imran & Almusharrarf, 2023; Steiss et al., 2024). Research into its role in assessment, however, remains insufficient.

To address this gap, this study compares the writing evaluation of pre-service teachers' (PTs' hereinafter) written texts conducted by teacher educators and ChatGPT. Three of us are teacher educators (TEs' hereinafter) who teach undergraduate courses on the theory and practice of writing education in universities. In a "Writing Education" course taught by one of us, 19 PTs were assigned to write an editorial for a university newspaper. After reviewing the PTs' written texts, we collaboratively developed evaluation criteria for the news editorial and rubrics to rate PTs' texts.

Building upon Kim et al. (2025, manuscript submitted for publication)—which compared the evaluation criteria and results developed by TEs and PTs for the same writing task—, this study uses the same dataset to extend the comparison to ChatGPT. The aim of this study is to explore the potential & limitations of generative artificial intelligence in writing evaluation and to discuss its implications for writing pedagogy.

Drawing on this background, this study examines how TEs and ChatGPT evaluate and interpret university newspaper editorials. In turn, we will examine what criteria and rubric TEs and ChatGPT develop for the editorials, and how TEs and ChatGPT rate the 19 texts based on the criteria by TEs. Accordingly, the research questions of this study are as follows:

1. What are the writing evaluation criteria developed by TEs and ChatGPT for newspaper editorials?
2. What are the rating results given by ChatGPT and the TEs based on the criteria developed by TEs?

II. Literature Review

1. Application of AI in General Education and Writing Education

In recent years, the emergence of generative artificial AI including ChatGPT has been transforming the overall modes of teaching, learning, and evaluation in education. Several studies have examined the functions, strengths, and limitations of ChatGPT and have argued that such technologies are expanding discussions about the future of learning, instruction, and evaluation in higher education(Allam et al., 2023; Rudolph et al., 2023). In particular, ChatGPT has been empirically tested within both teacher and learner contexts, and its educational applicability has been increasingly explored(Allam et al., 2023).

Although GPT-based tools have been utilized in various forms within higher education, few studies have systematically examined their effects at both micro and macro levels. To address this gap, Tajik(2024) analyzed the potential capabilities of GPT in three domains—student-facing, teacher-facing, and system-facing contexts. They have demonstrated that AI can support a wide range of educational activities in higher education. Meanwhile Mao, Chen, & Liu(2023) has noted that while generative AI holds potential for enhancing educational practices, it also leads to critical challenges and dilemmas when applied to teaching, learning, and evaluation.

On the other hand, Smolansky, Cram, Radulescu, Zeivots, Huber, & Kizilcec(2023) questioned whether traditional evaluation practices remain sustainable amid the rapid proliferation of generative AI. A university-based survey revealed that both instructors and students were already using AI to a certain extent, though their perceptions of its impact varied depending on the type of evaluation. Instructors tended to favor evaluation methods that assume AI use, whereas students expressed concerns about the possible decline in creativity. Such findings suggest that AI should not merely be treated as a

supplementary tool, but that evaluation design and implementation must be restructured in alignment with AI-integrated learning environments (Smolansky et al., 2023). This line of inquiry has direct implications for writing education, where evaluation plays a central role in shaping both learning processes and instructional practices.

Building on this discussion, generative AI has also driven significant changes beyond general education into the domain of writing instruction. Large language model (LLM)-based AI systems such as ChatGPT have gained attention as tools that support students' writing processes and supplement teachers' feedback and evaluation. Previous studies have shown that ChatGPT goes beyond simple information transfer and plays a role in providing multifaceted assistance throughout the entire writing process (AlAfnan et al., 2023; Imran & Almusharraf, 2023).

At the stages of exploring information and generating ideas, ChatGPT has been found to offer students accurate and reliable information. It implies that ChatGPT has potential as an alternative to conventional search engines (AlAfnan et al., 2023). Learners can rapidly access background knowledge related to a topic, outline of writing, and develop logical structure of text through ChatGPT, which serve as cognitive scaffolds during the early stages of writing. ChatGPT's responses go beyond factual explanations; they support learners' organization of ideas and recognition of issues, thereby increasing learners' engagement and efficiency in writing tasks (Imran & Almusharraf, 2023).

From the perspective of formative feedback, ChatGPT can serve as an alternative or supplementary tool in situations where immediate teacher intervention is difficult. Steiss et al. (2024) compared feedback provided by humans and ChatGPT and found that ChatGPT can deliver a certain level of criteria-based feedback that is useful for students in checking and revising their drafts. Although human feedback was found to be superior in qualitative depth, ChatGPT can still function as a practical support tool in large-scale writing classes or in contexts where timely teacher feedback is limited.

The potential of generative AI has also been highlighted in terms of linguistic quality and logical coherence. Herbold, Hautli-Janisz, Heuer, Kikteva, & Trautsch (2023) reported that argumentative texts produced by ChatGPT were rated higher than those written by humans in terms of logical structure and linguistic complexity. This finding suggests that generative AI can maintain a consistency in linguistic quality and produce texts with logical flow and refined expression.

ChatGPT has been applied to creative writing and genre-specific writing. Shidiq(2023) noted that while ChatGPT can support students' thinking and expressive abilities in creative writing, it also poses a potential risk of diminishing learners' creative capacities. Megawati(2023) examined the applicability of ChatGPT in the context of scientific writing, suggesting that AI can function as a writing assistant in articulating complex scientific concepts with greater clarity.

Comparative study between AI-generated and human-written texts have provided new directions for writing evaluation. Yan, Fauss, Hao, & Cui(2023) identified structural and linguistic differences between AI's writing and human's writing through large-scale textual analysis, and emphasized the need for methodological improvements to ensure the quality of writing evaluation. These findings indicate that AI has the potential to reshape the overall methodology of writing education and writing evaluation.

2. Applications of AI in Writing Evaluation

While ChatGPT has demonstrated potential across teaching & learning domains, recent research has increasingly explored its role in the field of writing evaluation. Writing evaluation is one of the most direct means of evaluating learners' thinking and expressive abilities, yet it has also been regarded as one of the most challenging areas due to the inherent subjectivity of raters and the heavy time demand. Human raters are likely to make divergent judgments on the same task, which can reduce the reliability and validity of the evaluation(Awidi,

2024; Wang, Engelhard Jr, Raczynski, Song, & Wolfe, 2017). Wang et al.(2017) reported that inter-rater consistency in writing evaluation was relatively low and that raters and experts differed in their perceptions about the degree of text borrowing, the development of ideas, and the maintenance of focus. These findings suggest that subjectivity is structurally embedded in human evaluation for written texts.

Awidi(2024) similarly pointed out that traditional human grading processes are inherently subjective, particularly when multiple raters are involved in large classes, often resulting in inconsistent scores. He noted that “writting assessment normally has subjective and inconsistent effects in grading,” emphasizing that inconsistent criteria and excessive workload among raters undermine both the consistency and efficiency of assessment. In educational contexts where one teacher is responsible for many classes and students, it is difficult to secure sufficient time for thorough evaluation and feedback, which may in turn lead to fluctuations in grading quality.

Given these practical constraints, the introduction of AI-based evaluation systems has been proposed as an alternative approach to enhance the objectivity and consistency of writing evaluation. Awidi(2024) raised the critical question, “Can AI tools like ChatGPT mark, assign a grade/score, and provide personalized feedback on reflective essays effectively?” He explored the potential of AI to evaluate open-ended writing tasks such as reflective essays in a rapid and consistent manner, arguing that generative AI tools like ChatGPT has the potential to complement human raters’ subjective judgments.

Jauhiainen and Guerra(2025) applied ChatGPT-4 to the evaluation of open-ended written examinations among university students to examine the extent to which AI scoring aligned with human grading. ChatGPT-4 evaluated 54 responses(ranging from 24 to 256 words) in a total of 3,240 times based on five criteria, showing a statistically significant correlation with human ratings. The researchers concluded that ChatGPT demonstrated promising potential in providing consistent and reproducible assessment of open-ended responses. This

finding suggests that ChatGPT can serve not merely as a text generator but as an assistant rater, enhancing the efficiency and reliability of the evaluation process.

In South Korea, similar efforts have emerged to incorporate ChatGPT into writing assessment. Park and Lee(2024) developed an AI-based writing-communication-competency- assessment-tool using ChatGPT, integrating three major categories—content, organization, and language use—into an automated scoring algorithm that provides both scores and domain-specific feedback. This study was significant in that it presented a concrete procedure for integrating generative AI into assessment frameworks. Similarly, Park and Lee(2024) compared rubrics for writing evaluation developed by human and AI, exploring the characteristics and educational implications of AI-generated rubrics. Their results indicated that while both human and AI raters prioritized content-related elements, AI introduced additional categories for criteria such as format, citation, critical thinking and overall impression. These differences imply that AI can serve to supplement or expand the human's evaluation framework.

Further, Korean studies have empirically examined the feasibility of AI-based automated scoring and feedback systems from multiple perspectives. Son(2023), Cha et al.(2024), and Choi et al.(2025) analyzed the potential of AI-assisted writing evaluation(AWE) systems and teacher–AI collaborative models, while Son(2025), Kim(2024), Lee(2025), and Choi(2025) explored the effectiveness and limitations of AI feedback tools in university writing instruction. These studies have shown that AI can improve the efficiency and consistency of assessment. However, most of them have focused primarily on automated scoring & feedback mechanism rather than on comparison between ChatGPT and human raters.

If we synthesize the results of previous studies, ChatGPT might have the potential to (1)mitigate raters' subjectivity, (2)promote the standardization of evaluation criteria, and (3)provide rapid feedback on many pieces of writing. Despite these possibilities, existing re-

search has not yet empirically verified whether ChatGPT can apply the same evaluation criteria as human raters or whether its scoring results are consistent and valid. This study compares the evaluation criteria developed by TEs and ChatGPT and analyzes what are the perspectives and frameworks for “good writing” by TEs and ChatGPT. Furthermore, this study examines how ChatGPT rates texts consistently based on the TEs’ criteria. Through this analysis, this study explores the level of validity and consistency exhibited by generative AI in both the process and outcomes of writing evaluation.

III. Research Method

To address ‘Research Questions 1 and 2’, this study proceeded through the following procedures, respectively.

First, one of us assigned a writing task, as shown in <Table 1>, to 19 undergraduate students(PTs). The participants were students majoring in Korean Language Education at the College of Education of K University, located in Gangwon Province, South Korea, who were enrolled in the course titled “ Writing Education”.

Table 1. Writing Task Assigned to Pre-service Teachers

※ Write an editorial for the university newspaper. The editorial should be approximately 1.5 pages (A4) in length and edited with the default settings of Hangul Word Processor (HWP). - An newspaper editorial is an opinion article that presents the author's perspective and argument on a specific issue in a newspaper or magazine. - Required features of an editorial: • Selection of a timely and relevant topic considering the reader(members of the university community) • Consistency of claims • Use of objective and sufficient grounds to support claims • Logical organization that appeals to readers' reasoning and empathy • Accurate, original, and persuasive expression • Appropriateness of title - Consider the media context in which the editorial will be communicated—although the article will appear in the printed newspaper, the title of the article will also be shown online to attract readers' clicks.
--

The task was designed to simulate authentic writing for the university newspaper's editorial section and to evaluate students' ability to construct persuasive arguments.

1. Procedure for Research Question 1

To address Research Question 1 three TEs collaboratively developed evaluation criteria and a rubric to rate 19 editorials written by PTs, and subsequently rated 19 texts. The TEs followed the steps summarized in <Table 2> for the development of criteria and rating.

Table 2. The Process of Development of criteria and Rating by TEs

Step 1. The three TEs jointly developed the evaluation criteria and rubric for rating through discussion and consensus, referring 19 texts written by PTs.
Step 2. Using the rubric, each TE independently rated sample texts(4 out of 19) based on the rubric from Step1.
Step 3. The three TEs discussed the results of the ratings from Step 2, and adjusted the criteria and rubrics accordingly.
Step 4. Each TE independently scored the remaining texts.
Step 5. The inter-rater reliability was calculated among the three TEs.
Step 6. The final score for each text was determined by averaging the three TEs' scores.

Before developing the rubric, the TEs carefully reviewed all 19 editorials. Then, due to all of us living in different cities, we had an online meeting. During the meeting we developed the rubric through discussion, taking into account both the characteristics of the university newspaper editorials and the quality levels of the PTs' writing. The online meeting was recorded for research purposes. After developing the rubric, the TEs first independently rated four texts. We then had a second online meeting to discuss again the criteria and rubric as well as discrepancies among our scores. Through discussion, we adjusted the criteria and rubric. Following this calibration session, each TE rated the remaining 15 texts independently.

Inter-rater reliability among the three raters was calculated using the intraclass correlation coefficient (ICC), which yielded $r = .969$

(95% CI = 0.934–0.987, $p < .001$), indicating a very high level of reliability. The final score for each text was determined by averaging the three raters' scores.

Next, we instructed ChatGPT to generate a set of writing evaluation criteria for a newspaper editorial and a corresponding rubric. This study employed OpenAI's ChatGPT Plus version(GPT-4 based), which provides enhanced data analysis and text-processing capabilities compared to the standard model(free subscription version). The Plus version was selected because it offered more stable performance for research purposes while remaining practically accessible for educational use.

ChatGPT was asked to develop its rubric under conditions equivalent to those of the TEs. To this end, the researchers allowed ChatGPT to review all 19 texts, and provided the same writing task(Table 1). No restrictions were imposed on the number of criteria, scoring range, or weighting of items, allowing ChatGPT to design its framework in its own way.

To ensure that ChatGPT's rubric achieved a level of specificity and detail comparable to that of TEs, the researchers continued to refine it's output through iterative prompting, issuing additional instructions after the initial dialogue(Table 3). Each round of prompting aimed to enhance the clarity, precision, and validity of the generated criteria.

Table 3. Example of Prompt–Response Sequence Used for Rubric Generation

Researcher:	I will assign you the role of a writing evaluator. The 19 texts I will provide are editorials written by pre-service teachers for the university newspaper. Based on the following conditions, review these texts and create a concrete set of evaluation criteria and a scoring rubric. (Writing Task was provided.)
ChatGPT:	Great! Based on the conditions you provided, I'll develop the assessment criteria and rubric. Please share the 19 essays so that I can analyze them and produce the detailed rubric accordingly.

2. Procedure for Research Question 2

To address Research Question 2 both the TEs and ChatGPT rated the same 19 texts based on the rubric created by the TEs. The rating process the TEs followed were the same as explained above (Table 2). The TEs' rubric and scoring guide was provided to ChatGPT. To ensure that ChatGPT's evaluation process resembled that of the TE's, it was first instructed to rate a pilot set of four texts. Analysis of the initial results showed that ChatGPT tended to assign higher scores than the human raters. Therefore, to ensure consistency in scoring rigor, we provided additional prompts to adjust the strictness within the same rubric framework—corresponding to the rater calibration phase in human evaluation. After re-evaluation, ChatGPT's score distribution aligned more closely with that of the TEs.

ChatGPT then evaluated all 19 texts, including the remaining 15. The range of scores given by ChatGPT was narrow. Accordingly, we additionally requested to apply the relative evaluation method to secure discrimination, and ChatGPT performed a re-evaluation accordingly. This adjustment was implemented to address the limited score range and to facilitate comparative analysis across texts. We instructed ChatGPT to specifically describe the grounds and reasons for scoring as well as the scores for each article, and the final results by ChatGPT were used as analysis data for the study. These explanations were used as key data for qualitative analysis.¹

All PTs' texts used in AI evaluation were fully anonymized, and no personally identifiable information was retained. During the use of ChatGPT, the data storage and history functions were disabled to prevent any exposure of personal information. All procedures complied

1 For Research Question 2, a consistent evaluation prompt was applied to all texts, assigning ChatGPT the role of an evaluator, specifying the evaluation context, providing the rubric, and requesting both scores and rationales. The full prompt log was not included in the main text due to space limitations.

with institutional research ethics standards.

3. Data Analysis

This study employed both content analysis and comparative analysis to examine the evaluation criteria and results by the TEs and ChatGPT. For Research Question 1, the two rubrics were subjected to content analysis. Each criterion in a rubric was examined in terms of its conceptual scope, evaluative focus (e.g., logic, grounds, expression), and linguistic specificity or inclusiveness. We independently reviewed and discussed our interpretations through iterative meetings until full consensus was reached.

For Research Question 2, this study conducted a descriptive comparison of the scores assigned to each category in the 'TEs' rubric. In addition, the qualitative content of the evaluative comments was compared to identify differences in focus (content-oriented vs. form-oriented), justification strategies, and feedback tone. These analyses were also reviewed independently by the three researchers and refined through repeated discussions to ensure consistency in interpretation.

IV. Results

1. Writing Evaluation Criteria & Rubric Developed by TEs and ChatGPT for a newspaper Editorial

First, the writing evaluation criteria and corresponding rubric developed by the TEs are presented below (Table 4). Each criterion consisted of three performance levels—Excellent (3 points), Average (2 points), and Inadequate (1 point)—with a total score of 33 points from 11 criteria.

Table 4. Criteria & Rubric Developed by TEs

Criterion 1. Topic Relevance	
1) Is the topic of the editorial timely and relevant to the current university context? 2) Does the topic present meaningful value or significance for discussion in a university newspaper editorial? 3) Is the topic chosen to appropriately consider the perspectives of university community members(students, faculty, and staff)?	
Excellent (3 points)	The editorial topic is timely and relevant to the current university context , addressing issues of public interest for the university community. The topic adequately considers various university members (students, faculty, staff).
Average (2 points)	The topic lacks timeliness or relevance to current university issues, or does not sufficiently consider diverse perspectives within the university community.
Inadequate (1 point)	The topic is neither timely nor relevant to current university issues and fails to consider multiple perspectives within the university community.
Criterion 2. Clarity of Claim	
1) Are the claims reasonable and clearly stated, showing possibility and feasibility for tackling the issue? 2) Are the claims original rather than obvious?	
Excellent (3 points)	The claims are reasonable, original, and clearly stated, that are possible and feasible to tackle an issue.
Average (2 points)	The claims are somewhat lack of possibility of solving an issue or feasibility. Or the claims are somewhat lacking in originality.
Inadequate (1 point)	The claims are not reasonable or vague, or overly obvious without analytical depth.
Criterion 3. Appropriateness of Argument	
1) Does the argument present meaningful and highly relevant issues, identifying the key points related to the topic? 2) Does the argument maintain logical and consistent claims in addressing the issues raised?	
Excellent (3 points)	The text presents meaningful & relevant issues, maintaining logical & consistent claims throughout the argument.
Average (2 points)	There is lack of consistency in maintaining arguments or lack of presenting relevant issues related to topic.
Inadequate (1 point)	The argument is inconsistent or unilaterally asserted or doesn't present relevant issues.
Criterion 4. Validity of Grounds	
1) Are the grounds relevant to the claims? 2) Do the grounds support the claims? 3) Are the grounds sufficient to justify the claims? 4) When source materials of grounds are presented, are they credible and properly cited?	

Excellent (3 points)	The grounds are relevant to the claims, sufficiently credible and supporting them. Sources are properly cited.
Average (2 points)	The grounds partially support the claims or source citation is incomplete.
Inadequate (1 point)	The grounds are insufficient, unrelated, or unreliable.

Criterion 5. Appropriateness of Title

Does the title accurately reflect the main idea and content of the editorial?

Excellent (3 points)	The title accurately reflects the main idea and content of the editorial.
Average (2 points)	The title is somewhat related to the content but does not clearly convey the main idea.
Inadequate (1 point)	The title is not connected to the content.

Criterion 6. Appropriateness of Organization

Does the text have an appropriate and complete structure with an introduction, body, and conclusion?

Excellent (3 points)	The texts have a clear and complete structure (introduction–body–conclusion).
Average (2 points)	The structure is partially complete but lacks appropriateness or balance.
Inadequate (1 point)	The structure is inappropriate or incomplete.

Criterion 7. Coherence

Is the text logically and coherently developed across the introduction, body, and conclusion?

Excellent (3 points)	The texts develops ideas logically and coherently across the introduction, body, and conclusion.
Average (2 points)	Some logical gaps or inconsistencies appear in the development of ideas.
Inadequate (1 point)	The text lacks coherence or includes repetitive or disorganized content.

Criterion 8. Connectivity of Sentences and Paragraphs

- 1) Are the connections between sentences natural and smooth in meaning?
- 2) Are the connections between paragraphs natural and coherent in meaning?

Excellent (3 points)	The connectivity of sentences and paragraphs are natural and smooth, with appropriate paragraph division.
Average (2 points)	Some connectivity of sentences or paragraphs are awkward or some paragraphing is inappropriate.
Inadequate (1 point)	There are many unnatural connections in meaning between sentences or between paragraphs, with inappropriate paragraph organization.

Criterion 9. Appropriateness of Vocabulary and Sentences

Does the text avoid inappropriate vocabulary, awkward expressions, literal translations, inanimate subjects, slang, and unnecessary abbreviations?

Excellent (3 points)	Uses appropriate vocabulary and sentences without using inappropriate words, awkward expressions, literal translations, inanimate subjects, slang, or unnecessary abbreviations.
Average (2 points)	One or two instances of inappropriate vocabulary, awkward expressions, literal translation, use of inanimate subjects, slang, or unnecessary abbreviations are found.
Inadequate (1 point)	More than three instances of inappropriate vocabulary, awkward expression, literal translation, inanimate subjects, slang and unnecessary abbreviations were found, reducing the appropriateness of the vocabulary and sentences.

Criterion 10. Grammatical Accuracy

- 1) Does the text follow correct spelling and spacing conventions?
 2) Is the text free from grammatical errors or awkward sentences, such as missing subjects, lack of subject–predicate corresponding, or ambiguous expressions?

Excellent (3 points)	No grammatical, spelling, or spacing or sentence structure errors ;
Average (2 points)	1–2 grammatical, spelling, or spacing errors or sentence structure errors are found.
Inadequate (1 point)	More than three grammatical, spelling, or spacing or sentence structure errors are found.

Criterion 11. Length Compliance

Is the text written in approximately 1,800 characters, excluding spaces?

Excellent (3 points)	Approximately 1,700–1,800 characters (excluding spaces).
Average (2 points)	Between 1,300 and 1,700 characters.
Inadequate (1 point)	Less than 1,300 characters.

The overall aim of the rubric was designed to reflect the characteristics of university newspaper editorials in detail, enabling a step-by-step evaluation from topical relevance to linguistic accuracy. The criteria & rubric generated by ChatGPT is presented in Table 5. The following table 6 presents the rubric generated by ChatGPT, based on eight criteria, with a total score of 100 points. Criteria 1–4 are rated on a 15-point scale each, while Criteria 5–8 are rated on a 10-point scale each. Each criterion accompanied by a detailed rubric describing performance levels across three bands—high, medium, and low.

Table 5. Criteria Generated by ChatGPT

Criterion 1. Appropriateness of Topic Selection • Is the topic relevant to the interests of university members? • Is it timely and worthy of discussion within the university community? • Does it contribute to the formation of public discourse rather than expressing a personal opinion?
Criterion 2. Consistency of Logical Development • Is the argument developed consistently and logically? • Are the claims presented in a coherent and natural flow? • Are supporting grounds systematically organized?
Criterion 3. Objectivity and Validity of grounds • Are the supporting grounds (examples, statistics, research findings, etc.) sufficient? • Are the sources credible and clearly identified? • Are the claims supported with concrete and valid evidence rather than mere assertions?
Criterion 4. Logical persuasiveness that resonates with the reader • Does the text persuade through logic rather than emotional appeal? • Does it analyze issues from the reader's perspective and appeal to their attention? • Does it anticipate counterarguments and address them effectively?
Criterion 5. Accuracy, Originality, and Persuasiveness of Expression • Are sentences clear and easy to understand? • Are there any original or impactful expressions? • Does the writer use neutral and persuasive language without emotional or biased tone?
Criterion 6. Appropriateness of Title • Does the title accurately reflect the content of the text? • Does it engage readers and encourage them to click? • Is it concise, specific, and intuitive rather than lengthy or vague?

Criterion 7. Media Appropriateness (Considering Newspaper and Online Contexts)

- Does the text conform to the editorial style (objective tone, logical structure, etc.)
 - Is it likely to attract readers' interest and be shared online?
 - Is it readable and engaging?
-

Criterion 8. Adherence to Length and Editorial Completeness

- Does the text comply with the required length (approximately 1.5 pages of A4)?
 - Is paragraph organization appropriate and are paragraphs smooth?
 - Are there no spelling, grammar, or sentence-structure errors?
-

Table 6. Rubric Generated by ChatGPT

1. Appropriateness of Topic Selection	
13–15 (Excellent)	Addresses a timely and socially significant topic relevant to the university community, contributing to public discourse. Demonstrates a distinctive and original awareness of the issue.
8–12 (Average)	The topic is meaningful to university students but lacks timeliness or has limited potential for broader public discussion. Remains at the level of familiar or conventional issues.
0–7 (Poor)	Focuses on a highly personal or irrelevant topic with little connection to university life. The issue awareness is unclear or of limited significance.
2. Consistency of Logical Development	
13–15 (Excellent)	The argument develops logically and coherently. The flow of ideas is natural and clear, with smooth paragraphing and a well-structured reasoning process.
8–12 (Average)	The overall logic is maintained, but some paragraphing is awkward or reasoning is somewhat inconsistent. The main argument may not always be clearly expressed.
0–7 (Poor)	The logic is disorganized and lacks consistency. The connection between claims and grounds is weak, and the flow between paragraphs is disjointed.
3. Objectivity and Validity of grounds	
13–15 (Excellent)	Effectively supports claims with diverse and credible sources (e.g., statistics, studies, examples). Sources are clearly identified and grounds are abundant.
8–12 (Average)	Grounds are presented but some sources lack credibility or clarity. The connection between claims and grounds is somewhat weak.
0–7 (Poor)	Lacks sufficient grounds or relies mainly on subjective opinions. Sources are unclear, unreliable, or distorted.
4. Logical persuasiveness that resonates with the reader	

13–15 (Excellent)	Analyzes the issue from the reader's perspective, building empathy through logical reasoning. Anticipates counterarguments and addresses them effectively. Balances objectivity with emotional appeal.
8–12 (Average)	Contains elements that readers can relate to, but logical persuasiveness is somewhat weak or counterarguments are insufficiently addressed. Consideration of readers is limited.
0–7 (Poor)	Presents one-sided claims without considering readers' perspectives. It is difficult to appeal to readers because of weak persuasiveness or illogical flaws.

5. Accuracy, Originality, and Persuasiveness of Expression

9–10 (Excellent)	Sentences are clear and highly readable. Expressions are original, and arguments are persuasive without relying on emotional appeals.
6–8 (Average)	Sentences are generally clear but contain occasional ambiguities. Lack of originality or weak persuasiveness.
0–5 (Poor)	Sentences are unclear and difficult to understand. Relies on emotional appeal or uses monotonous, unpersuasive expressions.

6. Appropriateness of Title

9–10 (Excellent)	The title effectively encapsulates the core content of the text while appealing to readers' attention with creativity.
6–8 (Average)	The title reflects the content but is somewhat ordinary or lacks intuitiveness. It lacks elements that capture the reader's interest.
0–5 (Poor)	The title does not match the theme of the text, or is overly long or vague, or not intuitive. Fails to attract readers' attention.

7. Media Appropriateness (Considering Print and Online Contexts)

9–10 (Excellent)	Well suited to the newspaper editorial format, with excellent readability. Uses organization and style effective for online dissemination.
6–8 (Average)	Generally appropriate for newspaper editorial format but somewhat less readable or engaging in online contexts.
0–5 (Poor)	Inconsistent with editorial conventions. Sentences are lengthy or complex, reducing readability; unlikely to attract attention online.

8. Adherence to Length and Editorial Completeness

9–10 (Excellent)	Maintains an appropriate length (approximately 1.5 A4 pages). Paragraph organization is smooth, with no spelling or grammatical errors.
6–8 (Average)	Slightly exceeds or falls short of the required length, but readability is not significantly affected. Spelling or grammatical errors exist.
0–5 (Poor)	The text is significantly shorter or longer than required. Paragraph organization is weak, and frequent language errors reduce readability.

A major similarity between the TEs and ChatGPT is that both sought to faithfully reflect the characteristics of a genre, newspaper editorial. Both recognized logical reasoning & persuasiveness as the essential foundation of an editorial, and incorporated this principle into their criteria.

The TEs emphasized content-oriented criteria such as “Topic Relevance”(timeliness and consideration of university members), “Clarity of Claim,” “Appropriateness of Argument,” and “Validity of Grounds,” alongside structure/meaning/language-related criteria including “Appropriateness of Organization,” “Coherence,” “Connectivity of Sentences and Paragraphs,” “Appropriateness of Vocabulary and Sentences,” and “Grammatical Accuracy.”

Similarly, ChatGPT identified “Consistency of Logical Development,” “Objectivity and Validity of Grounds,” and “Logical Persuasiveness” as core elements of evaluation. These parallels indicate that both evaluator groups(TEs and ChatGPT) understood the fundamental principles of persuasive writing, centering their judgments on logic and persuasiveness in text construction.

Another commonality lay in the establishment of a explicit and concrete grading criteria. The TEs adopted a three-point scale(high-medium-low), while ChatGPT used a 10 or 15-point scale with corresponding rubric bands. These detailed descriptions can allow reducing subjective variation and enhancing consistency and reliability among evaluators. In particular, the fact that ChatGPT, despite being an automated evaluation model, has constructed detailed evaluation criteria similar to those used by human evaluators suggests that AI may reflect certain characteristics of human assessment beyond the level of a simple automated scoring algorithm. However, such similarity does not imply that ChatGPT replicates human evaluators’ value systems or cognitive processes. Rather, ChatGPT appears to approximate human reasoning by inferring evaluative judgments from statistical patterns of language use.

On the other hand, there were clear differences in the compo-

sition and focus of the evaluation categories. First, the TEs' rubric contained 11 categories, while ChatGPT's rubric consisted of 8 more integrated categories. This difference reflects distinct perspectives on evaluation: the TEs designed a fine-grained rubric to examine student writing in detail and to provide targeted feedback, whereas ChatGPT produced key components into broader categories.

Second, in topic selection, the TEs emphasized "timeliness" and "consideration of university members." ChatGPT's criteria also incorporated these elements but further added "contribution to public discourse." In other words, ChatGPT conceptualized a university newspaper editorial not merely as an internal institutional text but as part of a larger social dialogue, expanding the scope of evaluation to include public relevance and potential for discourse formation. This aspect was rather overlooked by the TEs, which implies ChatGPT's criteria and rubric can be supplementary tools or references for cross-check.

Third, regarding logical development, the TEs treated "clarity of claims," "appropriateness of argument," and "appropriateness of organization" as separate categories, assessing how clearly claims were presented, how claims were connected to grounds and how the text's structure was logically formed. ChatGPT, by contrast, consolidated these aspects under the single category of "consistency of logical development," focusing on the overall coherence among argument, claims, and grounds as well as paraphrasing. Thus, while the TEs approached logic analytically through detailed components, ChatGPT used a criterion for overall logic coherence. ChatGPT evaluates 'counterarguments' as a part of "Logical persuasiveness that resonates with the reader " but we are doubtful that considering counterarguments is separated from judging 'logical development.' TEs think that when human evaluator rate a piece of persuasive writing, s/he can discern ①author's claims, ②grounds(reasons and evidences) for the claims, and ③logical building up of arguments. We could discern these 3 features in 19 texts so rate them differently according to quality of each

writing. The ChatGPT in this study didn't or couldn't perform this like human raters.

Fourth, the two groups diverged in their treatment of 'persuasiveness' and 'considering readers'. The TEs did not assign persuasiveness as an independent criterion, viewing it instead as a quality embedded within other categories such as appropriateness of argument, clarity of claims, and validity of grounds. ChatGPT, on the other hand, established "Logical persuasiveness that resonates with the reader" as a distinct category, including the ability to anticipate and respond to counterarguments. ChatGPT dealt with persuasiveness as part of a logical process as well as an aspect related to readers. This may be caused by the ChatGPT's info-processing of writing tasks "Logical organization that appeals to readers' reasoning and empathy." In contrast, the TEs viewed persuasiveness as an integrated effect—emerging from logical content-organization and coherence.

Fifth, in the domain of expression, the TEs separated "appropriateness of vocabulary and sentence use" from "grammatical accuracy," focusing on suitable use of words & sentences, and grammatical correctness. ChatGPT combined these dimensions into two categories: "accuracy, originality, and persuasiveness of expression," and "adherence to length and editorial completeness." In the former, it deals with clarity of sentences & persuasive expression, and in the latter, it checks grammatical correctness. ChatGPT didn't mention about the use of vocabulary and dealt with paragraphing as a part of completeness of text in "adherence to length and editorial completeness." TEs are well aware that students often make mistakes when writing Korean sentences, such as using ungrammatical sentences and inappropriate vocabulary. Therefore, they have incorporated sentences and vocabulary into separate assessment criteria. Since ChatGPT is more adept at processing information in English than in Korean, this aspect is believed to have been overlooked.

Sixth, the two groups differed in "originality." The TEs emphasized originality in thoughts or claims—the novelty of ideas and per-

spectives—whereas ChatGPT valued creativity in language use, such as expression.

Seventh, clear differences were also observed in terms of media appropriateness. While the TEs' rubric did not include a specific category for media format, ChatGPT introduced "media appropriateness" as an independent item, assessing both adherence to the conventions of newspaper editorials and readability in online environments. This demonstrates that ChatGPT viewed the "university newspaper editorial" not solely as a printed text but as one that circulates within digital media contexts. It seems that ChatGPT reflects the requirement of writing tasks into the criteria and rubric thoroughly. This aspect was disregarded in TEs' criteria.

Furthermore, in the "appropriateness of title" category, ChatGPT evaluated not only whether the title reflected the text's content but also whether it could attract readers' attention and encourage clicks. This approach shows that ChatGPT expanded the notion of writing evaluation beyond textual construction to include reader response and the digital environment. How such evaluation criteria should be interpreted in educational contexts remains open to further discussion. The TEs, by contrast, designed their rubric primarily for print-based editorials, focusing on representative feature of title rather than attracting readers' clicks.

The TEs' criteria are more detailed than ChatGPT's but the latter presents points in its criteria & rubric that the TEs had overlooked. In this regard, GPT could serve as a meta-cognitive tool for assessment that educators examine and recalibrate their own practice in setting criteria and developing rubrics. For example, educators may refer to the evaluation elements suggested by ChatGPT to examine whether certain criteria have been omitted or overly condensed in their existing rubrics, and to revise or adjust them accordingly.

2. Comparison Between TEs' and ChatGPT's Evaluation Results

Based on the rubric developed by the TEs, both ChatGPT and the TEs evaluated the same set of editorials written for the university newspaper. Each evaluation was conducted on a 33-point scale, and the comparison of scores and rankings is summarized as follows.

Table 7. Comparison of Scores between TEs and ChatGPT

Rank	TEs' Evaluation (Average)	Rank	ChatGPT's Evaluation (Total)
1	Text 13 (32 pts)	1	Text 9 (29 pts)
1	Text 18 (32 pts)	1	Text 4 (29 pts)
3	Text 16 (30.6 pts)	3	Text 3 (28 pts)
4	Text 7 (30 pts)	4	Text 7 (27 pts)
4	Text 8 (30 pts)	4	Text 13 (27 pts)
6	Text 10 (29.6 pts)	4	Text 1 (27 pts)
6	Text 19 (29.6 pts)	4	Text 17 (27 pts)
8	Text 17 (29.3 pts)	8	Text 6 (26 pts)
9	Text 6 (26.3 pts)	8	Text 2 (26 pts)
10	Text 4 (24.6 pts)	8	Text 11 (26 pts)
11	Text 5 (23.6 pts)	8	Text 14 (26 pts)
11	Text 9 (23.6 pts)	8	Text 16 (26 pts)
11	Text 11 (23.6 pts)	13	Text 8 (25 pts)
14	Text 2 (23.3 pts)	13	Text 12 (25 pts)
15	Text 14 (21.3 pts)	13	Text 5 (25 pts)
16	Text 3 (21 pts)	13	Text 15 (25 pts)
17	Text 15 (20.3 pts)	17	Text 18 (24 pts)
18	Text 1 (19 pts)	17	Text 10 (24 pts)
19	Text 12 (18.6 pts)	19	Text 19 (23 pts)

Although the two groups shared some overlap in overall score range, they showed marked differences in score distribution and evaluative focus. To verify the correlation between the evaluation scores of TE and ChatGPT, Spearman's rank correlation analysis was performed, and no significant relationship was found between the two variables ($\rho = -0.158$, $p = .52$, $N = 19$). This indicates that the TEs and ChatGPT applied distinct perspectives and priorities in judging the quality of the editorials.

First, ChatGPT exhibited a convergent rating tendency. The score distribution shows that ChatGPT clearly exhibited a central tendency bias. The TEs' scores ranged from a high of 32(13 and 18) to a low of 18.6 (12), a spread of about 13 points, indicating strong discrimination among texts. In contrast, ChatGPT's scores ranged only from 29(9 and 4) to 23(19), a spread of merely 6 points. This suggests that ChatGPT tended to avoid extreme scores and gravitated toward mid-range values.

This pattern indicates that ChatGPT tends to rely on mid-range scores when distinctions between levels are unclear. As a result, while its scoring was internally consistent, its capacity to discriminate subtle qualitative differences among texts was limited. In addition to this mid-range tendency, it is noteworthy that ChatGPT did not assign any text to the "Poor" band of its rubric. Because none of the student editorials were evaluated at the lowest performance level, the lower end of the scale remained unused, further narrowing the overall spread of ChatGPT's scores.

Second, clear differences in evaluation focus emerged in the assessment of high-ranking and low-ranking writings. The TEs rated 13 and 18 the highest, while ChatGPT gave top scores to 9 and 4. The TEs prioritized the public relevance of the topic, logical coherence, and depth of argumentation, whereas ChatGPT favored clarity of expression, structural completeness, and surface-level consistency. A similar pattern appeared in the lower score range. The TEs penalized deficiencies in reasoning and insufficient grounds more heav-

ily, while ChatGPT assigned moderate scores to texts that maintained clear structure and sentence-level accuracy despite weak argument. For example, 12, which received the lowest score from the TEs, was rated at a mid-level(25 points) by ChatGPT.

We interpret that these two tendencies are closely related. ChatGPT's narrow score range reflects its tendency to rely on clearly identifiable linguistic and structural features—such as grammar, sentence organization, and paragraph structure—that can be applied consistently across different texts. In contrast, the TEs' wider distribution stems from their interpretive engagement with each text's ideas and argumentative depth, which allowed for finer distinctions in quality of the texts. This suggests that differences in scoring patterns are not merely quantitative but also reveal that human evaluators demonstrate the ability to analyze and judge a text with a sense of wholeness in detail.

These findings highlight a clear tendency in the two evaluation approaches(or processing). It seems that ChatGPT focused primarily on surface-level form and structural coherence, identifying grammatical or structural issues. Conversely, TEs emphasized the sophistication of argumentaion, the social implications of the topic/theme and awareness of readers, which are the dimensions involving interpretive understanding, along with accuracy in expression and well-structured form.

V. Discussion

This study reviewed the writing evaluation by TEs and ChatGPT for university newspaper editorial, in order to explore the potential role of generative AI in writing assessment. The findings suggested that (1) GPT could function as a meta-cognitive reference that allows educators to examine and recalibrate their own practices in setting criteria and rubrics, and (2) ChatGPT applied the teacher-developed

rubric with a reasonable degree of fidelity, showing a tendency toward consistency across categories and reduced scoring variance. The teacher educators, on the other hand, demonstrated greater sensitivity to qualitative dimensions such as logical depth, social context, and reader awareness. These results suggest that while ChatGPT may be used to support consistency in writing assessment, human evaluators remain essential for contextual and interpretive judgment. The evaluation criteria developed by teacher educators can be understood as a synthesis of multiple professional interpretations, whereas AI-generated criteria represent the outcome of linguistic processing based on the input texts and instructions provided by the researcher.

That said, several risks must be considered when incorporating ChatGPT into writing assessment. Although AI systems are highly sensitive to linguistic form and surface completeness, they may fail to adequately capture the semantic depth, social implications, or rhetorical persuasiveness of a text. Shidiq(2023) cautioned that AI-based writing tools may simplify cognitive processes and weaken learners' critical thinking skills. Similarly, Tajik(2024) warned that generative AI carries risks of factual inaccuracy, bias, and plagiarism. These concerns equally apply to AI-driven or AI-supportive assessment: excessive reliance on automated judgment can undermine the reliability and ethical legitimacy of evaluation. Therefore, rather than accepting AI-generated results at face value, it is essential to include a human verification process in which teachers critically review and, when necessary, revise AI evaluations.

ChatGPT may be pedagogically utilized in writing assessment, particularly as a reference during the development of evaluation criteria. The rubrics generated by ChatGPT allow educators to examine their existing criteria for potential redundancy, omission, or imbalance. As discussed earlier, the differences between TEs and ChatGPT in criteria and rubric construction can prompt teachers to reflect on and reconsider their own assessment criteria. ChatGPT may be used to check the clarity and balance of evaluation criteria as a supplemen-

tary reference. In practical contexts, particularly when teachers conduct writing assessment independently, AI may serve as a supplementary reference for cross-checking evaluative judgments rather than as a substitute for human evaluation.

However, this study is limited by the small sample size, as it examined the evaluations of three teacher educators and 19 texts, which calls for caution in generalizing the findings. Nevertheless, the study is meaningful in that it makes explicit the differences in evaluative perspectives between teacher educators and generative AI in rubric-based writing assessment. While this study focused on rubric-based evaluation, future research could further investigate the practical feasibility of employing ChatGPT for consistency monitoring or formative feedback support. There is also a need to discuss the detrimental aspects of using generative AI for education, if there is any.

* Submitted 2025.11.26.
First revision received 2025.12.03.
Accepted 2025.12.18.

REFERENCES

- AlAfnan, M. A., Dishari, S., Jovic, M., & Lomidze, K. (2023). Chatgpt as an educational tool: Opportunities, challenges, and recommendations for communication, business writing, and composition courses. *Journal of Artificial Intelligence and Technology* 3(2), 60-68.
- Allam, H., Dempere, J., Akre, V., Parakash, D., Mazher, N., & Ahamed, J. (2023, May). Artificial intelligence in education: an argument of Chat-GPT use in education. In *2023 9th International Conference on Information Technology Trends (ITT)* (pp. 151-156). IEEE.
- Awidi, I. T. (2024). Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool. *Computers and Education: Artificial Intelligence* 6, 100226.
- Cha, J. W., Choi, S. H., & Lee, G. S. (2024). A Comparative Study of AI-Based Writing Assessment Tools and Human Evaluators – Focusing on book report texts –, *Korean Language and Literature in International Context*, (102), 681-705.
- Choi, J. Y., Kim, J. S., & Kim, H. S. (2025). Calibrating human and machine: Developing an AI-assisted rater discussion model for automated writing evaluation(AWE). *Journal of Korean Association for Educational Information and Media* 31(4), 1297-1329.
- Choi, S. G. (2025). Development and Validation of Quality Assessment Tools for AI-Powered Automated Writing Feedback, *Journal of CheongRam Korean Language Education*, (103), 227-263.
- Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific reports* 13(1), 18617.
- Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology* 15(4), ep464.
- Jang, S. M. (2023). ChatGPT has Changed the Future of Writing Education – Focusing on the response of writing education in the era of artificial intelligence –, *Research on Writing*, (56), 7-34.
- Jauhiainen, J. S., & Garagorry Guerra, A. (2025). Generative AI in education: Chat-GPT-4 in evaluating students' written responses. *Innovations in Education and Teaching International* 62(4), 1377-1394.
- Kim, J. Y. (2024). The Effects and Limitations of Using AI-Based Automatic Writing

- Feedback Programs in Writing Classes – Focusing on the Case of Using KEEwi in Reading & Expression Classes –. *The Journal of Yeongju Language & Literature* 58, 355-386.
- Kolade, O., Owoseni, A., & Egbetokun, A. (2024). Is AI changing learning and assessment as we know it? Evidence from a ChatGPT experiment and a conceptual framework. *Heliyon* 10(4).
- Lee, Y. B. (2025a). A Comparative Study on the Characteristics and Validity of Instructor, Peer, and AI Feedback in College Writing Education : Focusing on Common Issues in Students' Argumentative Texts. *Korean Journal of General Education* 19(3), 19-33.
- Lee, Y. B. (2025b). Reframing College Writing Education in the Era of Generative AI – An Integrated 'AI-Assisted Collaborative Writing Pedagogy' Combining Process-Based and Genre-Based Instruction with AI Literacy. *The Korean Journal of Literacy Research* 16(4), 175-207.
- Mao, J., Chen, B., & Liu, J. C. (2024). Generative artificial intelligence in education and its implications for assessment. *TechTrends* 68(1), 58-66.
- Megawati, R., Listiani, H., Pranoto, N. W., & Akobiarek, M. (2023). The role of GPT chat in writing scientific articles: A systematic literature review. *Jurnal Penelitian Pendidikan IPA* 9(11), 1078-1084.
- Park, S. Y., & Lee, B. Y. (2024). Exploring the Potential of AI Tools in University Writing Assessment: Comparing Evaluation Criteria between Humans and Generative AI. *Journal of practical engineering education* 16(5), 663-676.
- Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of applied learning and teaching* 6(1), 342-363.
- Shidiq, M. (2023, May). The use of artificial intelligence-based chat-gpt and its challenges for the world of education; from the viewpoint of the development of creative writing skills. In *Proceeding of international conference on education, society and humanity* (Vol. 1, No. 1, pp. 353-357).
- Smolansky, A., Cram, A., Radulescu, C., Zeivots, S., Huber, E., & Kizilcec, R. F. (2023, July). Educator and student perspectives on the impact of generative AI on assessments in higher education. In *Proceedings of the tenth ACM conference on Learning@ Scale* (pp. 378-382).
- Son, H. J. (2023). Perceptions and attitudes of English learners towards the utilization of AI-based Automated Writing Evaluation (AWE) programs and instructor feedback. *English Language & Literature Teaching*, 29(3), 91-112.
- Son, H. S. (2025). Studying the Potential Use of Generative AI in College Writing. *The*

- Journal of General Education*, (31), 41 - 72.
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., ... & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, 101894.
- Tajik, E. (2024). A comprehensive Examination of the potential application of Chat GPT in Higher Education Institutions.
- Wang, J., Engelhard Jr, G., Raczynski, K., Song, T., & Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing* 33, 36-47.
- Yan, D., Fauss, M., Hao, J., & Cui, W. (2023). Detection of AI-generated essays in writing assessments. *Psychological Test and Assessment Modeling* 65(1), 125 - 144.

교사교육자와 생성형 AI의 쓰기 평가 비교 연구

— 대학 신문 사설을 중심으로

안상희 · 김정화 · 김영란

본 연구는 예비교사의 글쓰기를 평가하는 과정에서 교사교육자(TEs)와 ChatGPT가 어떠한 평가 기준을 구성하여 채점하는지를 대학 신문 사설을 중심으로 탐구하였다. 이 연구에는 국어교육 전공 예비교사 19명이 참여하였다. 교사교육자 3인은 협의를 통해 평가 기준과 루브릭을 개발하여 텍스트를 평가하였으며, ChatGPT는 자체적으로 평가 기준과 루브릭을 생성한 뒤, 교사교육자가 개발한 루브릭을 적용하여 텍스트를 평가하였다.

교사교육자와 ChatGPT는 모두 사설 장르의 핵심 요소로 논리성과 설득력을 중시하였다. 그러나 교사교육자는 논증의 깊이와 독자 고려를 강조하는 내용 중심의 세분화된 평가 기준을 구성한 반면, ChatGPT는 명확성, 구조, 매체 적응성을 강조하는 통합적이고 형식 중심의 평가 기준을 제시하였다. 동일한 루브릭을 적용하여 채점한 결과, 교사교육자는 질적 민감성이 반영된 폭넓은 점수 분포를 보인 반면, ChatGPT는 중간 점수대에 수렴하는 점수 분포 양상을 보였다.

이러한 결과는 ChatGPT가 교사교육자가 개발한 루브릭을 비교적 일관되게 적용할 수 있으나, 맥락 고려나 해석에 따른 세밀한 의미 변화를 충분히 반영하는 데에는 한계가 있음을 시사한다. 또한 ChatGPT의 채점 결과는 점수의 변별력이 낮은 경향을 보였다. 그럼에도 불구하고 ChatGPT는 교사가 평가 기준과 루브릭 설계를 점검하거나 자신의 평가 실천을 검토할 때 보조적 도구로 활용될 가능성이 있어, 이는 글쓰기 평가에서 인간과 생성형 AI 간 협업 가능성을 시사한다.

핵심어 AI 활용 쓰기 평가, 교사 교육자, 생성형 AI, 챗지피티, 루브릭 개발

ABSTRACT

How Teacher-educators and ChatGPT Evaluate Pre-service Teachers' Writing?

— Focused on Newspaper Editorials

Ahn, Sanghee · Kim, Kyoungghwa · Kim, Youngran

This study explores how teacher educators (TEs) and ChatGPT construct and apply evaluation criteria in assessing pre-service teachers' (PTs) writing, focusing on university newspaper editorials. Nineteen PTs majoring in Korean language education participated in the study. Three TEs collaboratively developed evaluation criteria and a rubric to rate the texts, while ChatGPT (Plus version) was instructed to create its own rubric and later assess the same texts using the TEs' rubric.

Both TEs and ChatGPT emphasized logical reasoning and persuasiveness as core elements of the editorial genre. However, TEs developed detailed, content-oriented criteria highlighting argumentative depth and reader awareness, whereas ChatGPT presented integrated, form-oriented criteria emphasizing clarity, structure, and media adaptability. When applying the same rubric, TEs exhibited a wider scoring range with greater qualitative sensitivity, while ChatGPT showed a convergent mid-range scoring pattern.

These results indicate that ChatGPT can apply teacher-developed rubrics consistently but lacks contextual and interpretive nuance. The scoring results of ChatGPT also showed low discriminatory power of the scores. It may serve as a supplementary tool that helps teachers refine criteria & rubric design or cross-check their practice of evaluation, suggesting potential for AI-human collaboration in writing evaluation.

KEYWORDS AI-assisted writing evaluation, teacher educators, generative AI, ChatGPT, rubric development