INDONESIAN JOURNAL OF APPLIED LINGUISTICS
Vol. 16, No. 1, May 2026, pp. 58-80

Available online at:
https://ijal.upi.edu/index.php/ijal/article/view/73
https://doi.org/10.17509/yy1gpe90

ChatGPT as a formative support tool in German A2 writing assessment: An exploratory study

Pepen Permana*, Irma Permatawati, Nur Muthmainah, and Novia Anjani Dewi
German Language Education Study Program, Faculty of Language and Literature Education, Universitas Pendidikan Indonesia, Bandung, West Java, Indonesia

ABSTRACT

This study evaluates the feasibility of using ChatGPT as an instrument for assessing German writing skills at the CEFR A2 level. Motivated by concerns about subjectivity and inconsistency in traditional assessment, the study adopts an exploratory approach to examine ChatGPT's operational role as an AI-assisted tool, specifically analyzing its scoring behavior and its alignment with human judgment. Adopting an exploratory perspective, this study investigates ChatGPT's operational role as an AI-assisted assessment tool by examining its scoring behavior and alignment with human judgment. A mixed-methods complementary design was employed: qualitative descriptive analysis traced ChatGPT's assessment process, while quantitative analysis, using several relevant statistical tests, examined the consistency of ChatGPT's assessment results and their agreement with those from human raters. Writing samples were collected from 76 undergraduate students enrolled in a German-language education program at a public university in Indonesia. The writing tasks and assessment rubrics were adapted from the Goethe-Zertifikat A2 Modelsatz. The results indicate that ChatGPT demonstrated moderate consistency over time, reflected in significant positive correlations and general alignment with human raters. In addition, no significant differences were found in the mean scores assigned by ChatGPT and human raters. However, exact agreement on specific scoring decisions remained limited, revealing divergences in how AI and humans weigh linguistic nuances. The findings suggest that while ChatGPT can function as a valuable supplementary tool for formative assessment, human expertise remains essential for capturing the interpretive complexities of writing performance.

Keywords: Artificial intelligence; ChatGPT; German language; language assessment; writing skills

Received:	**Revised:**
21 July 2025	9 January 2026
Accepted:	**Published:**
4 May 2026	27 May 2026

How to cite (in APA style):
Permana, P., Permatawati, I., Muthmainah, N., & Dewi, N. A. (2026). ChatGPT as a formative support tool in German A2 writing assessment: An exploratory study. *Indonesian Journal of Applied Linguistics, 16*(1), 58-80. https://doi.org/10.17509/yy1gpe90

INTRODUCTION

The ability to write effectively in a foreign language, such as German, is essential because it enables learners to express complex ideas and supports effective communication (Saksono, 2022; Schnoor & Usanova, 2023). According to the Common European Framework of Reference for Languages (CEFR), the A2 level is an important stage where learners begin to be able to communicate in everyday situations with limited but functional language skills (Council of Europe, 2020; Li et al., 2024; North, 2021). Writing tasks at the A2 level involve producing longer, coherent texts, so assessing writing skills at this stage is important for evaluating learners' progress and language acquisition according to established standards. At this level, students are expected to be able to write simple, connected texts on topics that are familiar or of personal interest. They can describe experiences and events, dreams, hopes, and aspirations, and briefly provide reasons and explanations for their opinions and plans (Khushik & Huhta, 2020; Ma et al., 2023).

*Corresponding author
Email: pepen@upi.edu

However, in practice, assessing foreign-language writing skills often faces various challenges, mainly due to the subjectivity of human raters. For example, raters may differ in how they weigh linguistic errors against overall message clarity or in how they interpret the appropriateness of register in short written texts. While such interpretive judgment is widely recognized as an essential and legitimate component of writing assessment (Sickinger et al., 2025), subjectivity can also lead to inconsistency and bias in the assessment process (Tempelaar et al., 2020; Valentine et al., 2021), particularly when raters differ in how they prioritize aspects such as linguistic accuracy, task fulfillment, or communicative effectiveness. Such variability may also reduce the objectivity and reliability of the assessment outcomes, especially under conditions of limited time, high workload, or insufficient rater calibration. Several empirical studies have shown that there are variations and inconsistencies in language proficiency assessments, and underscored the necessity of finding solutions to ensure fairness and reliability in language assessment (Basile et al., 2022; Seo et al., 2025; Shabani & Panahi, 2020; Shirazi, 2019). These challenges are commonly observed in conventional human-rated writing assessments used in classroom-based evaluation and formal language testing contexts.

In addition, using traditional assessment methods for foreign language writing assignments and/or tests can create time constraints and limit the availability of qualified assessors. These limitations often result in a high workload for assessors and increase the potential for rating errors (Coombe et al., 2020; Kozlowska et al., 2019). Moreover, traditional assessment methods are time-consuming and require significant resources, which may hinder the effectiveness of language teaching and learning (Karataş et al., 2024; Ly, 2024; Song & Song, 2023).

Several studies have examined the development and use of automated writing evaluation (AWE) systems, such as e-rater and WriteToLearn, particularly in the context of English as a Foreign Language (EFL). The results show that compared to human assessment (Al-Inbari & Al-Wasy, 2023; Astutik et al., 2024; Geng & Razali, 2022). Due to their ability to provide instant feedback, these tools have received fairly good recognition (Huawei & Aryadoust, 2023). However, evidence from the literature also points to important limitations that complicate the direct transfer of these findings across assessment contexts. While reliability has often been emphasized, questions about the validity of automated assessments remain, particularly when AWE systems are applied to languages with structural characteristics that differ substantially from those of English. Previous studies have shown that AWE tools tend to perform less effectively in morphologically rich languages (Huawei & Aryadoust, 2023; Miranty & Widiati, 2021), such as German, which has features like case marking and verb position that can be challenging.

In response to these limitations, artificial intelligence (AI), especially large language models such as ChatGPT, has increasingly been examined in the literature. Unlike rule-based AWE tools, ChatGPT relies on large-scale language modeling (Gozali et al., 2024; Zhang, 2024). This architectural difference has been suggested to allow ChatGPT to adapt more flexibly to linguistic complexity (Fredrick & Craven, 2025; Shi et al., 2025). Accordingly, AI-assisted writing assessment has been discussed as a potential means of addressing certain sources of variability that often occur in human assessment. In addition, AI-based systems have been claimed to offer more consistent and measurable evaluations, though such claims require empirical verification (AlMakinah et al., 2025; Kondra et al., 2025).

Among the various foreign languages that can be assessed using AI, German is a relevant case study. This is because German has a unique grammatical structure, including inflectional morphology, case systems, and distinct noun genders. These features often pose challenges not only in the learning process but also in assessing writing skills (Kürschner & Nübling, 2011). In contexts such as Indonesia, where German is a relatively rarely taught foreign language, writing assessment is often conducted under conditions of limited instructional time, constrained access to trained assessors, and high teaching workloads. These contextual factors may affect how consistently and efficiently writing assessments are implemented in classroom settings. Therefore, examining how ChatGPT is used as a supportive assessment instrument in German writing skills, especially at the CEFR A2 level, is a relevant and important step. This exploration aims to understand the feasibility and adaptability of AI-based assessment tools within linguistically complex and resource-constrained educational environments.

In this context, the application of artificial intelligence (AI) technology, such as ChatGPT, appears to be a potential means of support in language education and assessment. ChatGPT is an AI-based language model designed to process and generate text across multiple languages, including German. This model uses deep learning techniques to generate text similar to human language based on the prompts it receives. This capability enables interactive and responsive engagement with written input, which has been discussed in the literature as potentially beneficial for learning-oriented feedback practices (Albadarin et al., 2024; Mahapatra, 2024; Minh, 2024). Previous studies have reported that ChatGPT can analyze written texts and generate feedback related to linguistic form and task

fulfillment (Baskara, 2023; Mahapatra, 2024). However, these abilities do not always yield valid, unbiased, or pedagogically appropriate assessment outcomes, especially when ChatGPT is used across different languages and educational contexts (Bulut et al., 2024; Urzúa et al., 2025). Therefore, the integration of ChatGPT into the assessment process needs to be carefully examined to ensure its potential to increase objectivity, reduce bias, and improve the efficiency of writing assessment in language learning is fully realized.

ChatGPT shows indeed promising potential in addressing the challenges of writing assessment, however, there is still a necessity to conduct an in-depth evaluation of its accuracy and consistency as a writing assessment tool. Several studies have highlighted the challenges in assessing writing skills and proposed AI technology as a solution (Ferrouhi, 2023; Hopfenbeck et al., 2023; Suárez et al., 2024), but most of these studies have focused on English language learning. As a result, the extent to which findings from English-based automated writing evaluation (AWE) research can inform assessment practices in other languages remains an open question.

Differences in linguistic structure, assessment demands, and pedagogical expectations suggest that results obtained in English may not be directly transferable to languages with substantially different grammatical characteristics. Studies examining ChatGPT's performance in morphologically rich languages such as German are limited, and some studies showed that generative AIs may struggle with specific linguistic features unless tailored to those contexts (Albeihi & Rice, 2025; Choudhury, 2023). Moreover, there is still a lack of research specifically exploring ChatGPT's assessment process in the context of A2-level German writing skills. This gap underscores the need for more exploration on how AI can conduct valid and reliable assessments of writing ability at the beginner level.

The novelty of this study lies not only in its research context but also in its conceptual and methodological focus. Rather than positioning ChatGPT as a general solution for writing assessment, this study examines its use as a formative support tool within a clearly defined assessment framework, with particular attention to assessment processes, consistency, and alignment with human judgement. Although AI-based assessment tools have been extensively investigated in the context of more frequently taught languages such as English, research on less commonly taught languages is still limited. Some recent studies have started to examine AI-assisted assessment and scoring in other languages, for example Spanish in AI-supported second language learning contexts (Huang & Cassany, 2025), as well as German and Mandarin using specialized automated scoring models (Kumari, 2024; Li, 2025). However, most of these studies tend to focus primarily on instructional support or surface-level performance outcomes rather than assessment validity and reliability. To bridge this gap, this study aims to provide empirical evidence on how ChatGPT performs on CEFR A2-level German writing tasks using an exploratory mixed-methods approach. Thus, this study is expected to contribute to the development of AI in German language education and teaching.

The increasing need for reliable, efficient, and measurable assessment tools in foreign language education highlight the urgency of this study. The lack of objectivity, consistency, and resource allocation of traditional assessment methods can hinder the creation of effective language teaching and learning processes. By focusing on exploring the feasibility of using ChatGPT as a German writing skills assessment tool, this study seeks to provide empirical insights into how AI-based assessment support may contribute to the assessment process, in the context of providing timely and consistent feedback to learners, and helping teachers manage the assessment workload more effectively. Moreover, as the demand for the creation of digital-based learning environments increases, this study contributes to discussions on innovative approaches in assessment that can be well integrated into online and hybrid teaching models. This study situates AI-assisted writing assessment within a formative classroom context, in which scoring and feedback are intended to support learners' writing development rather than to inform certification or high-stakes decisions.The main objective of this study is to examine the feasibility of using ChatGPT as an AI-assisted assessment support tool for A2-level German writing skills based on the CEFR. This study adopts an exploratory approach to investigate how ChatGPT performs when applied to writing assessment tasks at the beginner level. Specifically, this study seeks to answer the following research questions (RQs).

1. How does ChatGPT perform in assessing German writing tasks at level A2 based on predetermined assessment criteria?
2. How consistent are the writing assessment results provided by ChatGPT when administered at different times?
3. How do ChatGPT's writing assessment results correlate with those of certified human raters?

METHOD

Research Design

This study used a mixed-method approach with a complementary design. It combined qualitative descriptive analysis to document and explain the assessment process performed by ChatGPT, and quantitative analysis with statistical tests to assess

the feasibility of using ChatGPT as a tool for evaluating German writing skills at the A2 level of the CEFR. The qualitative component was intended to provide a process-oriented description of how ChatGPT applied the predefined assessment rubric and generated scores and feedback, rather than to conduct an interpretive or theory-driven qualitative analysis. The quantitative component addressed issues of consistency and agreement.

This study focused on the following three main research questions: (1) RQ1 exploring how ChatGPT performs writing assessments through qualitative descriptive analysis, (2) RQ2 investigating the consistency of ChatGPT's writing assessment results over time, and (3) RQ3 comparing ChatGPT's writing assessment results with those of certified human raters.

Participants

The participants in this study were 76 students enrolled in an undergraduate German language education program at a public university in Indonesia, consisting of 17 males and 59 females. They were third-semester students and had completed German language courses at the A1 and A2 levels. Most of them had no prior experience with German before entering the study program. Prior to data collection, participants were informed about the study's purpose and provided consent for their writing samples to be used for research and to be assessed by ChatGPT and two certified human raters.

Instruments

The instrument used in this study was a writing task in the form of *Freies Schreiben* (free writing), which is commonly used in the *Goethe-Zertifikat A2* test, a German language proficiency test administered by the Goethe-Institut to assess German language skills at various internationally recognized A2 levels. This writing assignment was adapted from the *Goethe Zertifikat A2 Modelsatz* (Goethe-Institut e.V., 2016).

This writing task consists of two sections. The first section asks participants to write a personal message in the context of maintaining friendships. The second section is a task to write a semi-official message that aims to organize an action. Each section is limited to 20–30 words.

The writing assessment criteria are also adapted from the *Goethe-Zertifikat A2 Modelsatz*. As shown in Figure 1, these criteria assess participants' writing across two main aspects: Task Fulfillment and Language. The Task Fulfillment aspect consists of two criteria: Language Function and Register. The Language aspect also has two criteria: Language Spectrum and Language Mastery.

Figure 1

Evaluation Criteria for German Writing Skills

Criterion				Score
Task Fulfillment*		Language		
Language Function	Register	Spectrum	Mastery	
All 3 language functions appropriate in content and extent	Situationally and partner appropriate	Appropriate and differentiated	Occasional mistakes do not impair understanding	5
2 language functions appropriate or 1 appropriate and 2 partly	Mostly situationally and partner appropriate	Mostly appropriate	Several mistakes do not impair understanding	3.5
2 language functions appropriate or 1 appropriate and 2 partly	Partly situationally and partner appropriate	Partly appropriate	Several mistakes partly impair understanding	2
1 language function appropriate or partly	No longer situationally and partner appropriate	Barely appropriate	Several mistakes significantly impair understanding	0.5
Text length less than 50% (10 words in Part 1; 15 words in Part 2) of the required word count or topic is off		Text consistently inappropriate		0

* Note: If the criterion "Task Fulfillment" is rated 0 points, the total score for this task is 0 points.

Regarding the Language Function, the assessment is based on the completeness and appropriateness of the language function in its content and scope. Higher scores indicate more appropriate and comprehensive use. The Register measures the appropriateness of the text to the situation and the conversation partner. The Language Spectrum focuses on the accuracy and variation in the use of vocabulary and grammatical structures. The Language Mastery measures the frequency and impact of errors on message comprehension. The score for each criterion ranges from 0 to 5 points, with reference to each available descriptor.

Research Procedures

To ensure that the evaluation process was conducted systematically, this study was carried out in several stages. The research procedure consisted of the following steps:

Task and Rubric Development

The writing tasks were adapted from the *Goethe-Zertifikat A2 Modelsatz* to match the A2 proficiency standards outlined in the CEFR framework. The writing assessment rubric used the standardized criteria from the *Goethe-Zertifikat A2 Modelsatz*, which covered four criteria: language function, register, language spectrum, and language mastery.

Data Collection

The participants were asked to complete a writing task, which was designed to assess their German writing proficiency at the A2 level. A total of 76 students participated by completing the A2-level German writing tasks in a supervised classroom setting. The participants responded on paper using handwritten answers. Each response was collected and manually transcribed verbatim into digital format to help with quick input during the evaluation process using ChatGPT. To enhance transparency, anonymized samples of students' written responses are provided in the Appendix.

Assessment Process

In this study, the assessment process was conducted within a formative classroom context. Although the rubric was adapted from the *Goethe-Zertifikat A2 Modelsatz*, it was used to support learning-oriented evaluation rather than certification or high-stakes decision making.

The writing results were first evaluated using ChatGPT-4. This version was chosen due to its improved performance and reliability compared to its earlier versions. It was prompted to act as a certified A2-level examiner and evaluate texts using the provided rubric. To address the known variability of generative AI outputs, the assessment was conducted using a fixed and consistent prompt, which functioned as an integral part of the assessment mechanism by constraining ChatGPT's evaluation to the predefined criteria.

The same set of writing samples was then independently assessed by two certified human raters using the same rubric. The raters hold the appropriate examiner license (*Prüferlizenz*) from Goethe-Institut for the relevant proficiency level. This means that the raters possess internationally recognized qualifications to assess German language skills according to CEFR standards.

Data Analysis

A qualitative descriptive approach was used to explore RQ1. This stage involved documenting and examining ChatGPT's writing assessment process through a rubric-guided analytical lens. It focused on how the model applied the predefined assessment criteria, interacted with input texts, generated feedback, and assigned scores. The qualitative analysis did not involve interpretive coding or thematic categorization, but rather a process-oriented examination aligned with the assessment rubric.

This analysis was based on direct observations of ChatGPT's evaluation interactions during the assessment sessions. It included how specific rubric criteria were operationalized in the generated feedback. For example, observations focused on whether ChatGPT explicitly identified the fulfillment of required language functions, commented on register appropriateness, or referred to grammatical accuracy when assigning scores. These observed patterns were then used to describe how the assessment process was carried out in relation to the predefined criteria. Quantitative data analysis was conducted to answer RQ2 and RQ3. This stage involved several statistical tests to measure the consistency of ChatGPT's writing assessments and their alignment with scores given by human raters. Pearson's Correlation Coefficient was used to measure the strength and direction of the linear relationship between scores from the first and second ChatGPT assessments. A strong correlation indicates that ChatGPT assessments are consistent over time.

Cohen's Kappa Coefficient evaluates the level of agreement between the two scores from the two rounds of ChatGPT's assessments and between ChatGPT and human raters. Higher Kappa values indicate stronger levels of agreement, reflecting the extent to which the assessments are aligned beyond the possibility of similarity occurring by chance (Li et al., 2023; Sim & Wright, 2005) Cronbach's Alpha was used to assess internal consistency, which is the extent to which a tool reliably assesses different aspects of the same skill over time (Anselmi et al., 2019). In this context, Cronbach's Alpha measures ChatGPT's consistency in applying assessment criteria to different writing tasks. Lastly, the paired t-test is used to compare the average scores given by ChatGPT and human raters. It determines whether there is a statistically significant difference between the average values of the two sets of scores.

FINDINGS

In this section, the research findings are presented in three stages according to the previously determined research questions (RQs). Overall, the findings reveal both patterns that support the feasibility of using ChatGPT for German A2-level writing assessment and indications of limitations that warrant cautious interpretation. First, the results of a qualitative descriptive analysis are presented to illustrate how ChatGPT performs the process of

assessing German writing tasks at level A2 (RQ1). Second, the results of a statistical analysis related to the consistency of ChatGPT's writing assessments are reported (RQ2). Third, the results of a statistical analysis comparing the writing assessment outcomes produced by ChatGPT and certified human raters are presented (RQ3)

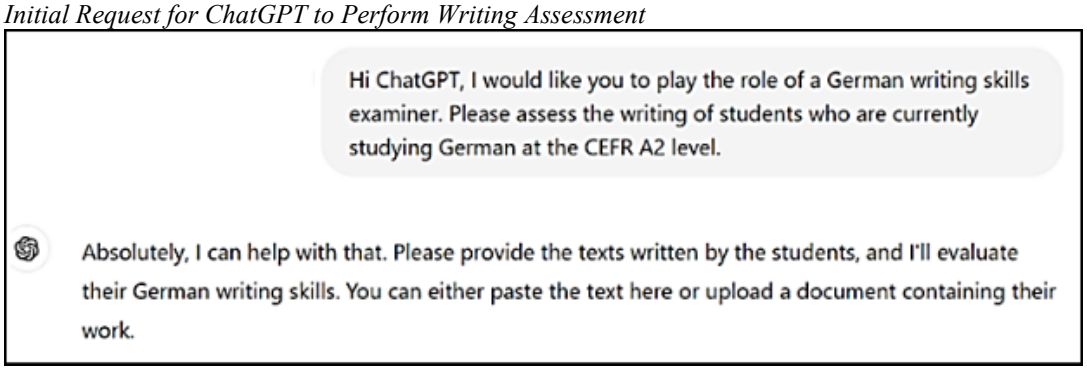
ChatGPT writing assessment process (RQ1)

This section addresses RQ1 by examining how ChatGPT operationalizes predefined CEFR A2 assessment criteria. Rather than providing a procedural walkthrough of the interaction, this analysis identifies analytical patterns in ChatGPT's evaluative behavior. Specifically, it examines how the model interprets rubric descriptors, the consistency of its linguistic evidence-gathering, and the logic used to synthesize scores and feedback. By categorizing these observable behaviors, this section provides empirical evidence of ChatGPT's internal "logic" when applied to the morphological and syntactic constraints of the German language.

The first step in this process, as seen in Figure 2, begins with a request from the user for ChatGPT to act as a German writing assessor. ChatGPT responds to this request by confirming this role. ChatGPT even directly asks the user to provide the student's text to be assessed. This stage is the starting point of the assessment process and establishes the conditions under which standardized assessment criteria can subsequently be applied.

Figure 2

Initial Request for ChatGPT to Perform Writing Assessment



This interaction shows that ChatGPT provides ease of interaction for its users. This allows for a smooth process of prompting through natural language processing. The user explicitly states that the assessment is for CEFR A2 level, which aims to ensure that the assessment is in accordance with the appropriate proficiency criteria. This explicit level specification constrains ChatGPT's evaluative scope and represents an initial alignment between the assessment request and the intended proficiency criteria, thereby shaping the relevance of the feedback generated.

The user then uploaded the assessment criteria and instructed ChatGPT to conduct an assessment based on those criteria. In response to the prompt, ChatGPT summarized the contents of the uploaded file by identifying the main components of the German A2-level writing assessment criteria. Although the uploaded criteria document was in German, ChatGPT was able to accurately and quickly summarize its content in English (see Figure 3). In this context, the user aimed to ensure that ChatGPT carried out the assessment process in accordance with the predetermined criteria, thereby establishing a shared evaluative framework prior to scoring. At this stage, ChatGPT's interpretation of the assessment rubric moved beyond a mere echo of the criteria; it was demonstrated through its ability to map specific linguistic markers in the students' texts onto distinct rubric dimensions. For instance, the model successfully distinguished between communicative task fulfillment (the presence of required information) and formal linguistic accuracy (such as German case endings and verb placement). By assigning specific textual evidence to these categories, ChatGPT operationalized the rubric as a granular analytical framework rather than a superficial checklist. This systematic mapping provided a standardized reference point that mitigated the subjective fluctuations often found in human rater' drift,' thereby supporting the reliability of the evaluative process in a formative context. Before proceeding to the assessment stage, ChatGPT first confirmed its understanding of the assessment format. As shown in Figure 4, ChatGPT stated its readiness to score the participants' writing based on two main categories: Task Completion and Language. In addition, ChatGPT confirmed the possible scores that could be assigned for each criterion. This confirmation stage was essential to maintain the accuracy of the assessment procedure. Thus, ChatGPT was expected to apply the appropriate assessment rubric when evaluating the participants' writing. This indicated that from the beginning of the assessment process, ChatGPT attempted to maintain the reliability and consistency of its evaluations.

Figure 3
Summary of Writing Assessment Criteria by ChatGPT

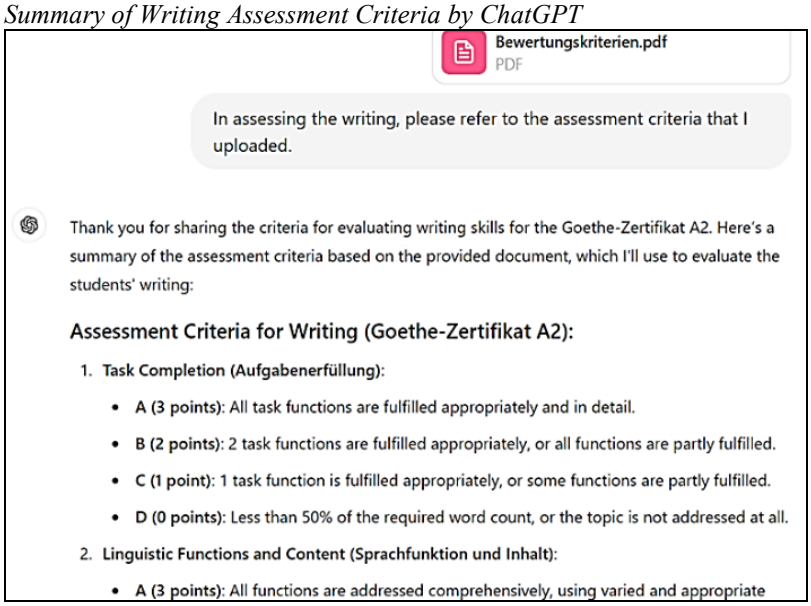
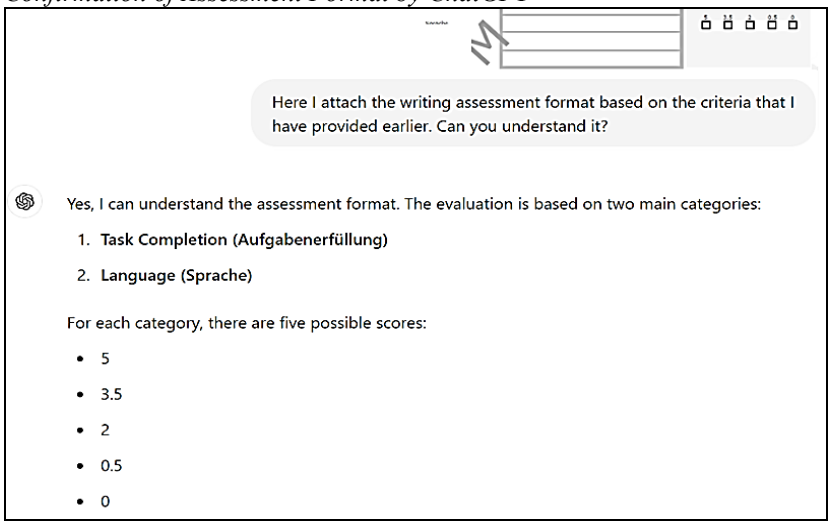


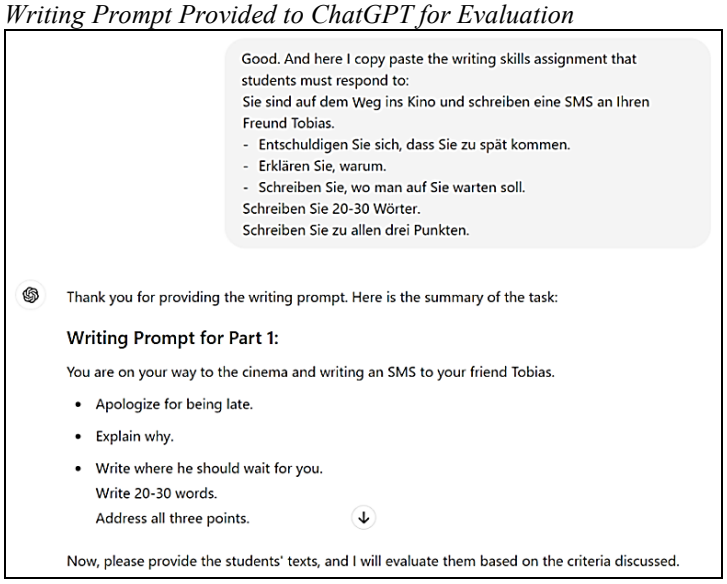
Figure 4
Confirmation of Assessment Format by ChatGPT



After the confirmation stage regarding the understanding of the assessment criteria, the user provided a prompt to ChatGPT containing the writing task instructions given to the participants. In this task shown in Figure 4, participants were asked to produce a short written message in a digital messaging context, which involved apologizing for being late, explaining the reason for the delay, and informing the meeting place. This task, one of the two writing tasks administered to the participants, was delivered in German and reflected a real-life communication situation relevant for CEFR A2-level learners. ChatGPT accepted the instruction and confirmed its readiness to assess the writing results according to the given task.

As seen in Figure 5, this stage indicates that ChatGPT identified and restated the key task requirements specified in the writing prompt and accurately summarized the three specific elements to be included in their writing. The task instructions stated that the participants' writing was limited to 20–30 words. This emphasized that the writing had to be clear and concise. With this well-defined task structure, the assessment process could be carried out in a focused manner. In addition, each participant's performance was evaluated based on comparable benchmarks. The design of the task also promoted the functional use of language and reflected everyday communication situations that might be relevant for the participants.

Figure 5
Writing Prompt Provided to ChatGPT for Evaluation



The next process was providing ChatGPT with the writing results of each participant. Based on the predefined rubric, ChatGPT evaluated each writing sample by applying the two main assessment aspects: Task Completion and Language, as illustrated in Figures 6a and 6b. The Task Completion aspect evaluates whether all task requirements were fulfilled. The Language aspect assesses the accuracy of vocabulary and grammar usage, as well as the coherence and fluency of the overall message. These two aspects are in line with the writing assessment standards of the *Goethe-Zertifikat A2*. This indicates that the evaluation was carried out using a standardized approach and in accordance with pedagogical principles.

Figure 6a
ChatGPT's Evaluation of a Student's German Writing Assignment

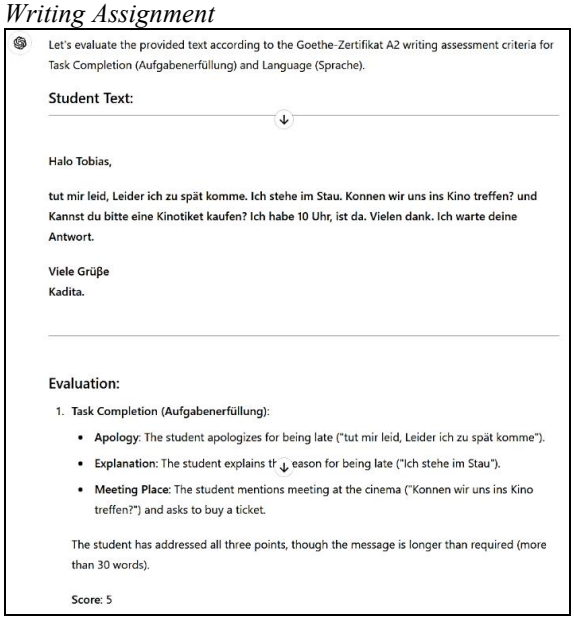
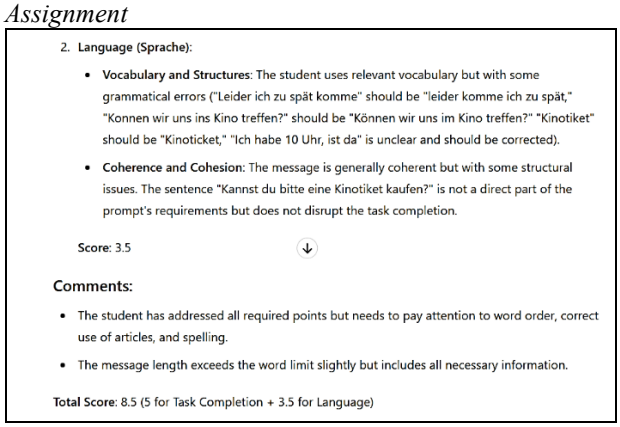


Figure 6b
ChatGPT's Evaluation of a Student's German Writing Assignment



During the assessment process, ChatGPT also generated explanatory feedback alongside score assignment. This feedback included explicit indications of whether the required task elements were fulfilled, as well as observations on grammatical accuracy, sentence structure, and vocabulary use. Specific mistakes in writing, such as incorrect word order, incorrect use of articles, and spelling errors, were also detected. Furthermore, ChatGPT provided suggestions for revision that were directly linked to the observed issues in the texts. In this way, the feedback extended beyond score reporting and functioned as formative input intended to support learners' awareness of strengths and areas for improvement.

Finally, ChatGPT was able to deliver a comprehensive evaluation, complete with a final score and detailed comments. As shown in Figure 6b, ChatGPT could assign a score to each aspect of the assessment based on how well participants had fulfilled the task requirements. ChatGPT was also able to offer suggestions for improvement, particularly regarding grammar and sentence coherence. For example, in the Language aspect, ChatGPT identified issues related to word order and grammatical form, such as *"Leider ich zu spät komme"*, which was suggested to be revised as *"Leider komme ich zu spät"* ("Unfortunately, I am late"), as well as article and verb usage, including *"Konnen wir uns im Kino treffen?"*, corrected to *"Können wir uns im Kino treffen?"* ("Can we meet at the cinema?").The detailed nature of this feedback, such as brief comments highlighting both strengths and areas for revision, explicitly identified which textual features influenced the assigned scores. Consequently, this can help participants gain a comprehensive understanding of the strengths in their writing as well as the areas that required further development. Within the formative assessment context of this study, such feedback accompanies score reporting and documents how evaluation criteria were applied, rather than functioning as a summative judgment for certification purposes.

Consistency of ChatGPT's writing assessments (RQ2)

To explore RQ2, which is related to the consistency of writing assessments by ChatGPT, the assessment of participants' writing results by ChatGPT was conducted twice at different times. Consistency in this study was examined through complementary statistical perspectives, recognizing that correlation alone does not fully capture stability in scoring decisions. Accordingly, three hypotheses were formulated: (1) There is a significant positive correlation between the writing assessments provided by ChatGPT at different times (H1); (2) There is substantial agreement between the writing assessments provided by ChatGPT at different times (H2); and (3) There is high internal consistency in ChatGPT's writing assessments over time (H3). These hypotheses were tested using several statistical analyses, including Pearson Correlation Coefficient, Cohen's Kappa, and Cronbach's Alpha with the help of SPSS version 27 software.

To test H1, the Pearson Correlation Coefficient was performed to measure the relationship between the assessments by ChatGPT at the first and second round. Based on the comparison of the total writing assessment scores, a correlation coefficient of 0.641 was obtained with a significance level of $p < 0.01$, as presented in Table 1.

Table 1

Correlation between 1st and 2nd evaluations by ChatGPT

Correlation		1st Evaluation	2nd Evaluation
1st Evaluation	Pearson Correlation	1	.641**
	Sig. (2-tailed)		.000
	N	70	70
2nd Evaluation	Pearson Correlation	.641**	1
	Sig. (2-tailed)	.000	
	N	70	70

Note: **. Correlation is significant at the 0.01 level (2-tailed).

This value indicates a statistically significant positive correlation and falls within the moderate range, following the interpretation proposed by Cohen (2013), where r values between 0.50 and 0.69 are considered moderate. This indicates that there is a fairly stable relationship between the assessment results given by ChatGPT at two different times. However, correlation reflects similarity in score patterns rather than exact agreement in individual scoring decisions. Accordingly, while these findings support H1 regarding temporal consistency at the score level, further analyses are required to examine stability in categorical assessment decisions.

To analyze H2, the Cohen's Kappa Coefficient was calculated to determine the level of agreement between the two rounds of assessment conducted by ChatGPT, focusing on consistency in categorical scoring decisions. As shown in Table 2, with a significance level of $p < 0.01$, the calculation obtained a Kappa value of 0.522.

Table 2

Measure of Agreement between 1st and 2nd Evaluations by ChatGPT

Measure of Agreement	Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Kappa	.522	.083	6.043	.000
N of Valid Cases	70			

Note: a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

This Kappa coefficient value indicates that there is a moderate level of agreement between the first and second rounds of ChatGPT's assessments. These results suggest that the Kappa value supports the formulation of H2. The moderate agreement indicates that ChatGPT demonstrated a reasonably

consistent application of scoring categories over time, although some differences between the two rounds remain. In other words, ChatGPT's evaluations are reasonably reliable, though not entirely identical over time. Moreover, the statistical significance of the findings indicates stability in assessment decisions at the categorical level, while also acknowledging that complete agreement across rounds was not achieved.

It is worthy to note that the calculation of Kappa does not assume the null hypothesis, which means that the analysis allows for the possibility of real agreement, not just agreement that might happen randomly (Li et al., 2023). However, the asymptotic standard error was used to assess the statistical significance of the result. The asymptotic standard error estimates how much the Kappa value might vary if the null hypothesis were true. This statistically significant Kappa value further supports H2, indicating that ChatGPT's assessments are fairly reliable.

To further evaluate the internal consistency of the ChatGPT writing assessment, H3 was tested using Cronbach's Alpha. As shown in Table 3, the obtained Alpha value of 0.648 falls within an acceptable range for exploratory research and indicates a moderate level of internal consistency across the assessment criteria.

Table 3
Cronbach's Alpha

Reliability Statistics	
Cronbach's Alpha	**N of Items**
.648	4

This result suggests that ChatGPT applied the assessment criteria in a generally consistent manner across writing samples, while also indicating variability in how individual criteria contributed to overall scores.

Comparison between ChatGPT and human raters (RQ3)

To evaluate RQ3, regarding the comparison between writing assessments by ChatGPT and certified human raters, three hypotheses were formulated: (1) there is a significant positive correlation between ChatGPT and human raters' assessments (H4); (2) there is a substantial level of agreement between ChatGPT and human raters' assessments (H5); and (3) there is no significant difference between the mean scores of ChatGPT and human raters (H6). These hypotheses were tested using Pearson's correlation coefficient, Cohen's Kappa, and a paired-sample t-test. Thus, the primary focus of this subsection is on statistical comparison.

The Pearson Correlation Coefficient calculation was conducted to test H4. This test aimed to determine the relationship between the scores given by ChatGPT and the scores given by human raters. As summarized in Table 4, the calculation resulted in the Pearson Correlation Coefficient value of 0.538, with a significance level of $p < 0.01$. This coefficient value means that between ChatGPT's scores and human raters' scores, there is a significant positive relationship. This coefficient also indicates a moderate strength of association. These results suggest similarity in overall scoring patterns rather than equivalence in individual assessment decisions. Accordingly, the results indicate that ChatGPT and human raters tend to produce comparable score distributions, while differences in evaluative judgment at the case level may still occur.

Table 4
Correlation between ChatGPT and Human Raters' Scores

Correlations		**ChatGPT**	**Human raters**
ChatGPT	Pearson Correlation	1	.538**
	Sig. (2-tailed)		.000
	N	70	70
Human raters	Pearson Correlation	.538**	1
	Sig. (2-tailed)	.000	
	N	70	70

Note: **. Correlation is significant at the 0.01 level (2-tailed).

After the correlation analysis, H5 was tested using Cohen's Kappa Coefficient. This test was used to assess agreement between ChatGPT's and human raters' scoring decisions at the categorical level. The results of the Cohen's Kappa Coefficient calculation, presented in Table 5, obtained a Kappa value of 0.183, with a significance level of $p < 0.01$. A Kappa value of this size is classified as low. This means that there is agreement between the ChatGPT assessment and the human rater, but the level of agreement is relatively low. This finding indicates that, despite similarities in overall score patterns, ChatGPT's individual scoring decisions did not consistently coincide with those of human raters. The contrast between the moderate correlation and the low Kappa value suggests that alignment at the aggregate score level does not necessarily translate into agreement on specific assessment decisions. However, the results of this statistically significant test confirm that the agreement between ChatGPT and human raters, although limited, is not due to chance.

Table 5
Measure of agreement between ChatGPT and human raters

Measure of Agreement	Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Kappa	.183	.069	2.616	.009
N of Valid Cases	70			

Note: a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

To evaluate H6, a paired sample t-test was conducted to compare the average scores given by ChatGPT and human raters. This test was performed to determine if the scores from ChatGPT significantly differed from human raters' scores. As detailed in Table 6, the t-test resulted in a modest average difference of 0.125 between the scores from ChatGPT and human raters, with a standard deviation of 2.761 and a standard error mean of 0.330. With a 95% confidence interval, the score difference is in the range of -0.533 to 0.783. This means that the actual average difference is likely to be in that range. The results indicate no statistically significant difference between the average scores produced by ChatGPT and those assigned by human raters. This finding suggests comparability at the aggregate score level, while not implying equivalence in individual scoring decisions or evaluative judgments.

Table 6
Paired samples test comparing ChatGPT and human raters' scores

Paired Differences	Mean	Std. Dev	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
				Lower	Upper			
Pair 1: ChatGPT - Human Raters	0.125	2.761	0.330	-0.533	0.783	0.379	69	

With a t-value of 0.379 with 69 degrees of freedom, a p-value of 0.706 was obtained. The p-value obtained is relatively high, and it can be interpreted that there is no statistically significant difference between the average scores given by ChatGPT and human raters. Thus, the findings support H6 by showing similarity in average scoring outcomes across the two assessment sources, without implying equivalence in individual evaluative judgment.

It is important to note that the comparison presented in this subsection is limited to quantitative score-based outcomes. While ChatGPT's assessments were accompanied by generated feedback as part of the evaluation output, the analysis of human raters in this study focused exclusively on their assigned scores, as systematic documentation of rater feedback was not part of the data collection. Consequently, the observed discrepancies between moderate score-level correlation, low categorical agreement, and non-significant mean differences should be interpreted in relation to scoring decisions rather than evaluative commentary. This clarification helps contextualize the statistical findings without extending the scope of analysis beyond the available data.

DISCUSSION

Understanding the ChatGPT Evaluation Process

One of the main contributions offered by this study is the explanation of how ChatGPT performed the assessment process. This approach is carried out to present the AI assessment mechanism clearly and transparently, to build trust among other AI users, such as teachers or students. This study identifies the affordances and constraints of AI-assisted assessment by providing a comprehensive analysis of how ChatGPT operationalizes predetermined rubrics. Rather than presenting a binary list of pros and cons, the findings interpret ChatGPT's performance in terms of evaluative strengths—such as procedural consistency and rapid evidence-mapping—and inherent limitations, including the potential for surface-level linguistic misinterpretations. This nuanced interpretation allows educators to discern which aspects of the assessment process are ripe for AI integration and which components necessitated by the CEFR A2 framework still require the expert interpretive judgment of a human rater.

The emphasis on the transparency process in this study refers specifically to the explicit documentation of the assessment process, including how ChatGPT interprets the assessment rubric, responds to task instructions, assigns scores, and provides written feedback. This focus aligns with previous research that emphasizes the value of explainability and interpretability in AI-based educational tools (Alamäki et al., 2024; Kastania, 2024; Perkins et al., 2024). These studies show that there is a significant increase in the level of acceptance and trust of educators towards AI technology if they understand how the AI system carries out evaluation (Al-Abdullatif, 2024; Zhang et al., 2025).

Understanding how the AI evaluation process works is particularly relevant when AI is considered

as part of instructional assessment contexts, where assessment outcomes are used to support ongoing learning rather than to replace high-stakes proficiency testing. In this study, the use of ChatGPT is situated within such a context, where AI-assisted assessment can help educators monitor students' language learning progress and identify areas that require further instructional attention. By making the assessment process explicit, educators may better judge when AI-generated evaluations can be relied upon and when human judgment remains necessary, especially for subjective or context-dependent aspects of writing (Gabelaia et al., 2024; Kılınç, 2024; Roe et al., 2025).

In this context, the integration of AI and human assessment can be understood as a pragmatic response to current challenges in foreign language education, particularly in less commonly taught languages such as German, where instructional resources and qualified assessors may be limited. Previous studies have suggested that hybrid assessment models can balance the efficiency of AI-supported data processing with the interpretive expertise of human raters, especially in handling creative and contextual language use (Passerini et al., 2025; Venigandla et al., 2024). The findings of this study support this perspective by indicating that while ChatGPT can assist in providing consistent and timely assessments, human involvement remains essential to ensure meaningful interpretation of student writing. Accordingly, AI functions most effectively as a complementary tool that supports, rather than replaces, human raters in the assessment process.

Consistency of ChatGPT's Writing Assessments

The significant correlation result ($r = 0.641$, $p < 0.01$) between the two rounds of ChatGPT assessments reflects that ChatGPT was able to provide consistent assessment results, even though they were performed at different times. This result is a significant finding because it proves that ChatGPT is able to maintain a stable level of reliability over time. This result is important because consistency in assessment is a fundamental factor in education to maintain the accuracy of monitoring student learning (Bayer et al., 2016).

Similar patterns of consistency have been reported in previous studies on AI-assisted assessment, although these studies were conducted in different contexts, such as large-scale educational measurement or technology-supported assessment systems beyond foreign language writing (Antunes & Hill, 2024; González-Calatayud et al., 2021). While these studies differ from the present research in terms of language focus, proficiency level, and assessment purpose, they collectively indicate that AI-based assessment tools can generate repeatable evaluation outcomes when applied under controlled conditions. In this sense, the correlation observed in this study aligns with the broader literature on AI-supported assessment, while remaining specific to the instructional and formative context of A2-level German writing assessment.

In addition to the correlation results, the moderate level of agreement reflected in the Cohen's Kappa value ($\kappa = 0.522$, $p < 0.01$) in this study also confirms that ChatGPT was able to produce reasonably stable assessment outcomes across the two rounds of evaluation. This Kappa value suggests a moderate level of agreement, while also pointing to the presence of variability in how individual writing performances were evaluated over time. Such variability may be associated with the challenges of consistently applying standardized criteria to diverse writing forms and content. As discussed by Elkhatat (2023) and Bašić et al. (2023), differences in the type of writing assignment, the complexity of the language used by students, or even nuances in how AI interprets various linguistic elements could be factors that influence this variability. However, similar issues are also common in human assessments, as noted by Fogal (2020) and Dockrell and Connelly (2021), who reported that maintaining consistency between raters and between tasks was difficult. Therefore, while the results of this study indicate that ChatGPT has the potential to be a reliable assessment instrument, continued development is always needed for AI to handle more subtle assessment dimensions accurately.

In practical terms for educators, the moderate levels of agreement observed in this study suggest that while ChatGPT can reliably support the assessment of several aspects of student writing, human supervision is still necessary, especially for subjective elements such as tone or creativity. Within instructional or formative assessment settings, such levels of agreement may still be pedagogically valuable, as AI-assisted assessment can help reduce teachers' workload while providing timely feedback to learners. However, the presence of variability in assessment decisions indicates that further refinement is required before such systems can be considered appropriate for high-stakes assessment contexts.

Furthermore, the moderate Cronbach's Alpha value ($\alpha = 0.648$) produced in this study suggests that ChatGPT has a reasonable level of internal consistency. However, there are still some aspects to be improved, particularly to maintain the uniform application of scoring criteria across different types of writing tasks. Cronbach's Alpha is widely used to measure internal consistency. Values above 0.6 are generally considered adequate for exploratory studies, but in critical assessment situations, higher values are strongly recommended. The Cronbach's Alpha values obtained in this study might reflect the complexity of the language assessment process, where even small differences in task difficulty or

language structure may impact the stability of the scores awarded (Kuiken, 2023; Robinson, 2001). Given that writing tasks in educational contexts can assess different skill dimensions (Chen, 2024), these findings highlight the necessity of continuous development and refinement of AI technology like ChatGPT to maintain reliability in assessing various writing tasks.

Comparative Analysis of ChatGPT and Human Raters Scoring Accuracy

The evaluation of ChatGPT assessment accuracy compared to human assessment resulted in a moderate positive correlation ($r = 0.538$, $p < 0.01$). This suggests that ChatGPT assessments are generally in alignment with human judgment, but it is not fully capable of replicating the subtle decision-making process of human raters. This finding is consistent with the literature, primarily conducted in the context of second language learning, showing that while AI-based tools are quite effective, they often face issues in evaluating subjective aspects of language assessment, such as creativity, style, and contextual relevance (Davoodifard & Eskin, 2024; Rahman et al., 2024; Zhao et al., 2024). While these studies differ from the present research in terms of language focus and instructional context, they provide a relevant conceptual reference for interpreting the observed discrepancies between AI-based and human assessments in this study.

The study also yielded a relatively low Cohen's Kappa value ($\kappa = 0.183$, $p < 0.01$). This value means that there is only a slight level of agreement between the scores given by ChatGPT and human raters. This result means that there are certain areas where AI-based evaluations can differ from human assessments. This low level of agreement is a common problem in the use of AI in writing assessment, as AI sometimes fails to capture contextual aspects or nuances of writing that are more easily identified by human raters (Clark et al., 2021; Zhao et al., 2024). In the context of this study, the low Kappa value highlights a critical limitation of relying solely on AI-generated assessments. While ChatGPT can produce structured and repeatable evaluations, the observed disagreement indicates that human supervision remains necessary to interpret context-dependent features of learner writing. This finding underscores that AI-assisted assessment should be approached as a complementary tool rather than an autonomous decision-making mechanism.

Interestingly, the paired sample t-test, which compared the mean score between ChatGPT and human raters, revealed no significant difference ($t(69) = 0.379$, $p = 0.70$). This result indicates that, on average, the overall scoring tendencies of ChatGPT and human raters were similar. This result is positive, as it serves as evidence that ChatGPT is capable of mirroring human judgment in writing assessments. However, the absence of a significant difference in mean scores should not be interpreted as evidence of high agreement at the level of individual assessments. In other words, the fact that the average scores between ChatGPT and certified human raters in this context did not differ does not mean that differences in certain aspects should be ignored.

When considered alongside the low Kappa value, the t-test result suggests that comparable average scores may coexist with substantial variation in how individual writing performances are evaluated. Previous studies have likewise noted that AI tools often show stronger alignment with human raters on more objective features, such as grammar and vocabulary, while remaining less reliable in assessing subjective or context-sensitive aspects of writing (Escalante et al., 2023; Zhao et al., 2023). Taken together, these findings reinforce the need to interpret AI–human alignment with caution and to situate AI-assisted assessment within a human-in-the-loop framework.

Practical Implications and Research Opportunities

Viewed from an instructional perspective, the findings of this study highlight the role of AI-assisted assessment as a formative support mechanism rather than a substitute for human judgment. The implications offered by this study are the opening of opportunities to integrate AI in the assessment process, in this case, the assessment of foreign language writing skills, particularly as a supportive tool within instructional contexts. This potential is based on the consistency shown by ChatGPT in providing assessment scores. It should also be noted that this study found low agreement in specific assessment decisions between ChatGPT and human assessors. The coexistence of stable overall scoring patterns and decision-level discrepancies indicates that AI-assisted assessment cannot function independently of human judgment. Rather, the findings suggest that collaboration between human assessors and AI systems is advisable, especially when assessment decisions require contextual interpretation. This perspective aligns with previous studies emphasizing the continued role of human involvement in ensuring fairness and interpretive accuracy in assessment practices (Bignami et al., 2023; Elkhatat, 2023; Xia et al., 2024).

One practical implication of these findings for language assessment methodology is that ChatGPT may be effectively positioned as a tool to support formative assessment practices within language courses. In such contexts, AI-assisted assessment can provide timely feedback and help reduce teachers' assessment workload, while assessment outcomes are used to inform instruction rather than

to make high-stakes decisions. When designing assessment strategies that incorporate AI, educators should consider the results of this research. The use of AI, such as ChatGPT, can indeed increase efficiency, but a human role is still important, especially in assessing complex and meaningful aspects of language. These aspects cannot yet be completely accommodated by AI (Bowman et al., 2022; Hagos et al., 2024; Park et al., 2024).

While the findings offer useful insights, the study is limited by its focus solely on A2-level writing tasks in one language and a specific participant group, which could potentially restrict the generalizability of the findings. In addition, the assessment in this study was conducted using rubric-based scoring without a fine-grained linguistic analysis of A2-level writing features. As a result, certain discrepancies observed between AI-based and human assessments could not be further explained in terms of specific linguistic phenomena.

Further research should explore the performance of ChatGPT across varied language levels, writing tasks, and learner profiles, while also incorporating more detailed linguistic profiling of learner writing. Such approaches could provide a stronger explanatory basis for understanding how AI systems evaluate language performance. Additionally, future studies are also needed to improve the AI model to be more sensitive in recognizing the nuances of writing. This improvement may be achieved through training on more diverse datasets or by implementing hybrid evaluation models that integrate both AI and human judgment (Islam et al., 2024).

This study contributes to the growing body of empirical research on the use of AI in education, particularly in the context of assessing German writing skills at the CEFR A2 level. The findings indicate that while ChatGPT can support writing assessment by providing structured and temporally consistent evaluations, human involvement remains essential to interpret assessment outcomes, especially in cases requiring contextual or nuanced judgment.

CONCLUSION

The main objective of this study was to examine the feasibility of using ChatGPT as an instrument for assessing German writing skills at CEFR level A2. The focus of this study was (1) to describe how the writing assessment process is carried out by ChatGPT, (2) to test the consistency of ChatGPT assessments carried out at two different times, and (3) to measure the accuracy and alignment of ChatGPT assessments compared to certified human raters. The results of the analysis demonstrated that ChatGPT was able to perform the process of assessing German writing skills at level A2 transparently and consistently and could provide formative feedback based on the given rubric criteria.

Notably, the results of the study also showed that ChatGPT demonstrated moderate consistency over time, as reflected in a significant positive correlation and a moderate level of agreement between repeated assessments. This is proven by the acquisition of a significant positive correlation value and a moderate level of agreement between the two sets of assessment results. Furthermore, when compared to the results of certified human raters, the results of the study showed a moderate correlation and no significant difference in the mean scores given by ChatGPT and human raters. However, the relatively low level of agreement in specific scoring decisions suggests that AI-based assessments do not fully replicate human judgment, particularly in more nuanced evaluative aspects.

The findings of this study may suggest that ChatGPT can function as a complementary tool for writing assessment in foreign language learning, particularly in formative contexts and in settings with limited assessment resources While ChatGPT may support assessment efficiency and procedural consistency, human raters remain essential for ensuring fairness and for interpreting qualitative and contextual dimensions of writing that extend beyond automated scoring.

ACKNOWLEDGEMENTS

This research was supported by the Indonesian Ministry of Education, Culture, Research, and Technology under the 2024 Fiscal Year Research Grant. The authors gratefully acknowledge this financial support.

AI USE DISCLOSURE

ChatGPT's role in the preparation of this article manuscript is solely for the purpose of checking the use of language and grammar. ChatGPT provided suggestions on the draft of this article regarding clarity and coherence throughout the text. This aimed to ensure that this manuscript has good readability and flow and meets academic English standards. Additionally, ChatGPT also assisted in providing suggestions on the flow of the introduction section so that it is presented with a logical and systematic structure.

REFERENCES

Al-Abdullatif, A. M. (2024). Modeling teachers' acceptance of generative artificial intelligence use in higher education: The role of AI literacy, intelligent TPACK, and perceived trust. *Education Sciences*, *14*(11), Article 11. https://doi.org/10.3390/educsci14111209

Alamäki, A., Khan, U. A., Kauttonen, J., & Schlögl, S. (2024). An experiment of AI-based assessment: Perspectives of learning preferences, benefits, intention, technology affinity, and trust. *Education Sciences*, *14*(12), Article 12. https://doi.org/10.3390/educsci14121386

Albadarin, Y., Saqr, M., Pope, N., & Tukiainen, M. (2024). A systematic literature review of empirical research on ChatGPT in education. *Discover Education*, *3*(1), Article 60. https://doi.org/10.1007/s44217-024-00138-2

Albeihi, H. H. M., & Rice, M. F. (2025). Generative AI and language diversity: Implications for teachers and learners. *Arab World English Journal, 16*(1), 43-54. https://dx.doi.org/10.24093/awej/vol16no1.3

Al-Inbari, F. A. Y., & Al-Wasy, B. Q. M. (2023). The impact of automated writing evaluation (AWE) on EFL learners' peer and self-editing. *Education and Information Technologies*, *28*(6), 6645–6665. https://doi.org/10.1007/s10639-022-11458-x

AlMakinah, R., Goodarzi, M., Tok, B., & Canbaz, M. A. (2025). Mapping artificial intelligence bias: A network-based framework for analysis and mitigation. *AI and Ethics*, *5*(3), 1995–2014. https://doi.org/10.1007/s43681-024-00609-0

Anselmi, P., Colledani, D., & Robusto, E. (2019). A comparison of classical and modern measures of internal consistency. *Frontiers in Psychology*, *10*, 2714. https://doi.org/10.3389/fpsyg.2019.02714

Antunes, B., & Hill, D. R. C. (2024). Reproducibility, replicability and repeatability: A survey of reproducible research with a focus on high performance computing. *Computer Science Review*, *53*, 100655. https://doi.org/10.1016/j.cosrev.2024.100655

Astutik, I., Widiati, U., Ratri, D. P., Jonathans, P. M., Nurkamilah, N., Devanti, Y. M., & Harfal, Z. (2024). Transformative practices: Integrating Automated Writing Evaluation in higher education writing classrooms - A systematic review. *Indonesian Journal on Learning and Advanced Education*, *6*(3), 423–441. https://doi.org/10.23917/ijolae.v6i3.23675

Bašić, Ž., Banovac, A., Kružić, I., & Jerković, I. (2023). ChatGPT-3.5 as writing assistance in students' essays. *Humanities and Social Sciences Communications*, *10*(1), Article 750. https://doi.org/10.1057/s41599-023-02269-7

Basile, V., Caselli, T., Balahur, A., & Ku, L.-W. (2022). Editorial: Bias, subjectivity and perspectives in Natural Language Processing. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.926435

Baskara, F. R. (2023). Integrating ChatGPT into EFL writing instruction: Benefits and challenges. *International Journal of Education and Learning*, *5*(1), Article 1. https://doi.org/10.31763/ijele.v5i1.858

Bayer, S., Klieme, E., & Jude, N. (2016). Assessment and evaluation in educational contexts. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing Contexts of Learning: An International Perspective* (pp. 469–488). Springer International Publishing. https://doi.org/10.1007/978-3-319-45357-6_19

Bignami, E., Montomoli, J., Bellini, V., & Cascella, M. (2023). Uncovering the power of synergy: A hybrid human–machine model for maximizing AI properties and human expertise. *Critical Care*, *27*(1), Article 330. https://doi.org/10.1186/s13054-023-04598-0

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olah, C., Amodei, Daniela, Amodei, Dario, Drain, D., Li, D., Tran-Johnson, E., … Kaplan, J. (2022). *Measuring progress on scalable oversight for Large Language Models* (arXiv:2211.03540). arXiv. https://doi.org/10.48550/arXiv.2211.03540

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C. M., Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S. N., Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. *Chinese/English Journal of Educational Measurement and Evaluation*, *5*(3), Article 3. https://doi.org/10.59863/MIQL7785

Chen, T.-H. (2024). The effects of task complexity on L2 English rapport-building language use and its relationship with paired speaking test task performance. *Applied Linguistics Review*, *15*(2), 737–769. https://doi.org/10.1515/applirev-2021-0199

Choudhury, M. (2023). Generative AI has a language problem. *Nature Human Behaviour*, *7*(11), 1802–1803. https://doi.org/10.1038/s41562-023-01716-4

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's `human' is not gold: Evaluating human evaluation of generated text. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 7282–7296). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.565

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Coombe, C., Vafadar, H., & Mohebbi, H. (2020). Language assessment literacy: What do we need to learn, unlearn, and relearn? *Language Testing in Asia*, *10*(1), Article3. https://doi.org/10.1186/s40468-020-00101-6

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Council of Europe Publishing. https://www.coe.int/en/web/common-european-framework-reference-languages/companion-volume

Davoodifard, M., & Eskin, D. (2024). Exploring the role of generative AI in second language education: Insights for instruction, learning, and assessment. *Studies in Applied Linguistics and TESOL*, *24*(1), Article 1. https://doi.org/10.52214/salt.v24i1.12863

Dockrell, J. E., & Connelly, V. (2021). Capturing the challenges in assessing writing: Development and writing dimensions. In T. Limpo & T. Olive (Eds.), *Executive Functions and Writing* (p. 103-136). Oxford University Press. https://doi.org/10.1093/oso/9780198863564.003.0005

Elkhatat, A. M. (2023). Evaluating the authenticity of ChatGPT responses: A study on text-matching capabilities. *International Journal for Educational Integrity*, *19*(1), Article 1. https://doi.org/10.1007/s40979-023-00137-0

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*(1), 57. https://doi.org/10.1186/s41239-023-00425-2

Ferrouhi, E. M. (2023). *Evaluating the accuracy of ChatGPT in scientific writing*. https://doi.org/10.21203/rs.3.rs-2899056/v1

Fogal, G. G. (2020). Investigating variability in L2 development: Extending a complexity theory perspective on L2 writing studies and authorial voice. *Applied Linguistics*, *41*(4), 575–600. https://doi.org/10.1093/applin/amz005

Fredrick, D. R., & Craven, L. (2025). Lexical diversity, syntactic complexity, and readability: A corpus-based analysis of ChatGPT and L2 student essays. *Frontiers in Education*, *10,* 1616935. https://doi.org/10.3389/feduc.2025.1616935

Gabelaia, I., Dolidze, T., & Vasadze, N. (2024). Artificial Intelligence and ChatGPT in language education: Improving EFL classroom assistance and changing assessment methods. In V. Dermol (Ed.), *Proceedings of the MakeLearn, TIIM & PIConf International Conference 2024* (pp. 147–156). ToKnowPress. https://doi.org/10.53615/978-961-6914-31-4

Geng, J., & Razali, A. (2022). Effectiveness of the automated writing evaluation program on improving undergraduates' writing performance. *English Language Teaching*, *15*(7), Article 7. https://doi.org/10.5539/elt.v15n7p49

Goethe-Institut e.V. (2016). *Goethe-Zertifikat. A2: Modellsatz Erwachsene*. Goethe-Institut.

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial Intelligence for Student Assessment: A Systematic Review. *Applied Sciences*, *11*(12), Article 12. https://doi.org/10.3390/app11125467

Gozali, I., Wijaya, A. R. T., Lie, A., Cahyono, B. Y., & Suryati, N. (2024). Leveraging the potential of ChatGPT as an automated writing evaluation (AWE) tool: Students' feedback literacy development and AWE tools integration framework. *The JALT CALL Journal*, *20*(1), Article 1. https://doi.org/10.29140/jaltcall.v20n1.1200

Hagos, D. H., Battle, R., & Rawat, D. B. (2024). *Recent advances in generative AI and Large Language Models: Current status, challenges, and perspectives* (arXiv:2407.14962). arXiv. https://doi.org/10.48550/arXiv.2407.14962

Hopfenbeck, A. T. N., Zhang, Z., Sun, S. Z., Robertson, P., & McGrane, J. A. (2023). Challenges and opportunities for classroom-based formative assessment and AI: A perspective article. *Frontiers in Education, 8*, 1270700. https://doi.org/10.3389/feduc.2023.1270700

Huang, S., & Cassany, D. (2025). Spanish language learning in the AI era: AI as a scaffolding tool. *Journal of China Computer-Assisted Language Learning*. https://doi.org/10.1515/jccall-2024-0026

Huawei, S., & Aryadoust, V. (2023). A systematic review of automated writing evaluation systems. *Education and Information Technologies*, *28*(1), 771–795. https://doi.org/10.1007/s10639-022-11200-7

Islam, Md. S., Kabir, M. N., Ghani, N. A., Zamli, K. Z., Zulkifli, N. S. A., Rahman, Md. M., & Moni, M. A. (2024). Challenges and future in deep learning for sentiment analysis: A comprehensive review and a proposed novel hybrid approach. *Artificial Intelligence Review*, *57*(3), Article 62. https://doi.org/10.1007/s10462-023-10651-9

Karataş, F., Abedi, F. Y., Ozek Gunyel, F., Karadeniz, D., & Kuzgun, Y. (2024). Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. *Education and Information Technologies, 29*, 19343-

19366 https://doi.org/10.1007/s10639-024-12574-6

Kastania, N. P. 'Paulina.' (2024). Building trust in AI education: Addressing transparency and ensuring trustworthiness. In D. Kourkoulou, A.-O. (Olnancy) Tzirides, B. Cope, & M. Kalantzis (Eds.), *Trust and inclusion in AI-mediated education: Where human learning meets learning machines* (pp. 73–90). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64487-0_4

Khushik, G. A., & Huhta, A. (2020). Investigating syntactic complexity in EFL learners' writing across Common European Framework of Reference levels A1, A2, and B1. *Applied Linguistics*, *41*(4), 506–532. https://doi.org/10.1093/applin/amy064

Kılınç, S. (2024). Comprehensive AI assessment framework: Enhancing educational evaluation with ethical AI integration. *Journal of Educational Technology and Online Learning*, *7*(4-ICETOL 2024 Special Issue), Article 4. https://doi.org/10.31681/jetol.1492695

Kondra, S., Medapati, S., Koripalli, M., Nandula, S. R. S. C., & Zink, J. Z. (2025). AI and Diversity, Equity, and Inclusion (DEI): Examining the potential for AI to mitigate bias and promote inclusive communication. *Journal of Artificial Intelligence and Machine Learning*, *3*(1), Article 1. https://doi.org/10.55124/jaim.v3i1.249

Kozlowska, A., Wisniewski, Z., & Goossens, R. H. M. (2019). Methods for assessing the effectiveness of language learning – a comparative study. *Advances in Social and Occupational Ergonomics*, 97–106. https://doi.org/10.1007/978-3-319-94000-7_10

Kuiken, F. (2023). Linguistic complexity in second language acquisition. *Linguistics Vanguard*, *9*(s1), 83–93. https://doi.org/10.1515/lingvan-2021-0112

Kumari, P. (2024). The impact of AI models on students' writing competence and writing motivation in the context of learning German as a foreign language in India. *IJFMR - International Journal For Multidisciplinary Research*, *6*(2). https://doi.org/10.36948/ijfmr.2024.v06i02.18516

Kürschner, S., & Nübling, D. (2011). *The interaction of gender and declension in Germanic languages*. *45*(2), 355–388. https://doi.org/10.1515/flin.2011.014

Li, M., Gao, Q., & Yu, T. (2023). Kappa statistic considerations in evaluating inter-rater reliability between two raters: Which, when and context matters. *BMC Cancer*, *23*(1), Article 799. https://doi.org/10.1186/s12885-023-11325-z

Li, Q. (2025). AI-Assisted Writing in L2 Mandarin: Evaluating Impact and Student Perceptions in UK GCSE Classrooms. *International Chinese Language Education Communications*, *2*(1), 69–100. https://doi.org/10.46451/iclec.20250512

Li, Z., Li, J. Z., Zhang, X., & Reynolds, B. L. (2024). Mastery of listening and reading vocabulary levels in relation to CEFR: Insights into student admissions and English as a medium of instruction. *Languages*, *9*(7), Article 7. https://doi.org/10.3390/languages9070239

Ly, H. H. (2024). A review of alternative assessment methods and how to apply them in EFL classrooms. *Journal of Research in Humanities and Social Science*, *12*(5), 176–181. https://questjournals.org/jrhss/papers/vol12-issue5/1205176181.pdf

Ma, H., Wang, J., & He, L. (2023). Linguistic features distinguishing students' writing ability aligned with CEFR levels. *Applied Linguistics, 45*(4),637–657. https://doi.org/10.1093/applin/amad054

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, *11*(1), Article 9. https://doi.org/10.1186/s40561-024-00295-9

Minh, A. N. (2024). Leveraging ChatGPT for enhancing English writing skills and critical thinking in university freshmen. *Journal of Knowledge Learning and Science Technology*, *3*(2), Article 2. https://doi.org/10.60087/jklst.vol3.n2.p62

Miranty, D., & Widiati, U. (2021). An automated writing evaluation (AWE) in higher education. *Pegem Journal of Education and Instruction, 11*(4), 126–137. https://doi.org/10.47750/pegegog.11.04.12

North, B. (2021). The CEFR Companion Volume—What's new and what might it imply for teaching/learning and for assessment? *CEFR Journal - Research and Practice*, *4*, 5–24. https://doi.org/10.37546/JALTSIG.CEFR4-1

Park, Y.-J., Pillai, A., Deng, J., Guo, E., Gupta, M., Paget, M., & Naugler, C. (2024). Assessing the research landscape and clinical utility of large language models: A scoping review. *BMC Medical Informatics and Decision Making*, *24*(1), Article 72. https://doi.org/10.1186/s12911-024-02459-6

Passerini, A., Gema, A., Minervini, P., Sayin, B., & Tentori, K. (2025). Fostering effective hybrid human-LLM reasoning and decision making. *Frontiers in Artificial Intelligence*, *7*, 1464690. https://doi.org/10.3389/frai.2024.1464690

Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment

Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice*, *21*(06), Article 06. https://doi.org/10.53761/q3azde36

Rahman, A., Raj, A., Tomy, P., & Hameed, M. S. (2024). A comprehensive bibliometric and content analysis of artificial intelligence in language learning: Tracing between the years 2017 and 2023. *Artificial Intelligence Review*, *57*(4), Article 107. https://doi.org/10.1007/s10462-023-10643-9

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27–57. https://doi.org/10.1093/applin/22.1.27

Roe, J., Furze, L., Perkins, M., & Kitson, C. (2025, March 17). *The AI assessment scale: A practical framework for TESOL educators in the age of ChatGPT*. TESOL Connections | TESOL International Association. https://www.tesol.org/the-ai-assessment-scale-a-practical-framework-for-tesol-educators-in-the-age-of-chatgpt/

Saksono, L. (2022). Write recount text learning using a genre-based approach in German Literature class. *IJORER: International Journal of Recent Educational Research*, *3*(4), 403–413. https://doi.org/10.46245/ijorer.v3i4.196

Schnoor, B., & Usanova, I. (2023). Multilingual writing development: Relationships between writing proficiencies in German, heritage language and English. *Reading and Writing*, *36*(3), 599–623. https://doi.org/10.1007/s11145-022-10276-4

Seo, H., Hwang, T., Jung, J., Kang, H., Namgoong, H., Lee, Y., & Jung, S. (2025). Large Language Models as evaluators in education: Verification of feedback consistency and accuracy. *Applied Sciences*, *15*(2), Article 2. https://doi.org/10.3390/app15020671

Shabani, E. A., & Panahi, J. (2020). Examining consistency among different rubrics for assessing writing. *Language Testing in Asia*, *10*(1), Article 12. https://doi.org/10.1186/s40468-020-00111-4

Shi, H., Chai, C. S., Zhou, S., & Aubrey, S. (2025). Comparing the effects of ChatGPT and automated writing evaluation on students' writing and ideal L2 writing self. *Computer Assisted Language Learning*, *39*(4), 1148–1175. https://doi.org/10.1080/09588221.2025.2454541

Shirazi, M. A. (2019). For a greater good: Bias analysis in writing assessment. *Sage Open*, *9*(1), 2158244018822377. https://doi.org/10.1177/2158244018822377

Sickinger, R., Brunfaut, T., & Pill, J. (2025). Comparative Judgement for evaluating young learners' EFL writing performances: Reliability and teacher perceptions of holistic and dimension-based judgements. *Language Testing*, *42*(2), 137–166. https://doi.org/10.1177/02655322241288847

Sim, J., & Wright, C. C. (2005). The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, *14*, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

Suárez, A., Díaz-Flores García, V., Algar, J., Gómez Sánchez, M., Llorente de Pedro, M., & Freire, Y. (2024). Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. *International Endodontic Journal*, *57*(1), 108–113. https://doi.org/10.1111/iej.13985

Tempelaar, D., Rienties, B., & Nguyen, Q. (2020). Subjective data, objective data and the role of bias in predictive modelling: Lessons from a dispositional learning analytics application. *PLOS ONE*, *15*(6), e0233977. https://doi.org/10.1371/journal.pone.0233977

Urzúa, C. A. C., Ranjan, R., Saavedra, E. E. M., Badilla-Quintana, M. G., Lepe-Martínez, N., & Philominraj, A. (2025). Effects of AI-Assisted Feedback via Generative Chat on Academic Writing in Higher Education Students: A Systematic Review of the Literature. *Education Sciences*, *15*(10), 1396. https://doi.org/10.3390/educsci15101396

Valentine, N., Durning, S., Shanahan, E. M., & Schuwirth, L. (2021). Fairness in human judgement in assessment: A hermeneutic literature review and conceptual framework. *Advances in Health Sciences Education*, *26*(2), 713–738. https://doi.org/10.1007/s10459-020-10002-1

Venigandla, K., Vemuri, Navya, & Vemuri, Naveen. (2024). Hybrid intelligence systems combining human expertise and AI/RPA for complex problem solving. *International Journal of Innovative Science and Research Technology (IJISRT), 9*(3), 2066–2075. https://doi.org/10.38124/ijisrt/IJISRT24MAR2039

Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher*

Education, *21*(1), Article 40. https://doi.org/10.1186/s41239-024-00468-z

Zhang, C., Hu, M., Wu, W., Kamran, F., & Wang, X. (2025). Unpacking perceived risks and AI trust influences pre-service teachers' AI acceptance: A structural equation modeling-based multi-group analysis. *Education and Information Technologies*, *30*(2), 2645–2672. https://doi.org/10.1007/s10639-024-12905-7

Zhang, Y. (2024). Incorporating ChatGPT as an automated written corrective feedback tool into L2 writing class. *Journal of Language Teaching*, *4*(4), Article 4. https://doi.org/10.54475/jlt.2024.024

Zhao, R., Zhuang, Y., Zou, D., Xie, Q., & Yu, P. L. H. (2023). AI-assisted automated scoring of picture-cued writing tasks for language assessment. *Education and Information Technologies*, *28*(6), 7031–7063. https://doi.org/10.1007/s10639-022-11473-y

Zhao, Y., Zhang, R., Li, W., Huang, D., Guo, J., Peng, S., Hao, Y., Wen, Y., Hu, X., Du, Z., Guo, Q., Li, L., & Chen, Y. (2024). Assessing and understanding creativity in large language models. *arXiv:2401.12491v1.* https://doi.org/10.48550/arXiv.2401.12491

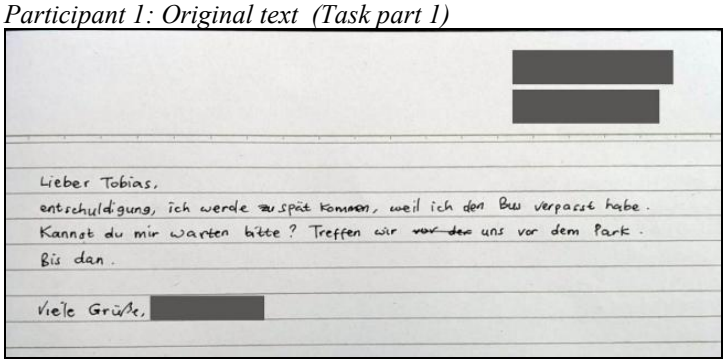
APPENDIX
Sample Participant Writing Task

Writing Task Part 1:

Sie sind auf dem Weg ins Kino und schreiben eine SMS an Ihren Freund Tobias.
- Entschuldigen Sie sich, dass Sie zu spät kommen.
- Erklären Sie, warum.
- Schreiben Sie, wo man auf Sie warten soll.

Schreiben Sie 20-30 Wörter.
Schreiben Sie zu allen drei Punkten.

Participant 1: Original text (Task part 1)



Participant 1: Transcribed text

Student Text:

Lieber Tobias,

entschuldigung, ich werde zu spät kommen, weil ich den Bus verpasst habe. Kannst du mir warten bitte? Treffen wir uns vor dem Park.
Bis dann.

Viele Grüße,

Participant 1: ChatGPT's evaluation

Evaluation:

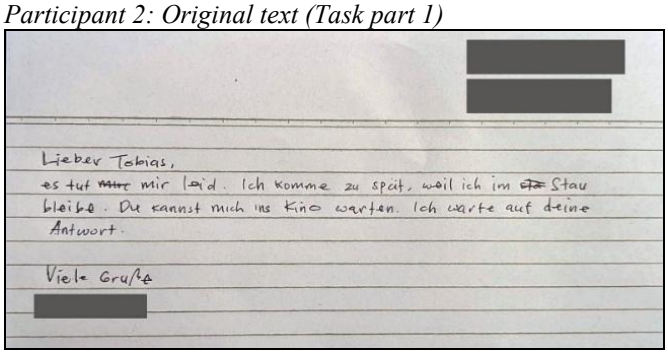
1. **Task Completion (Aufgabenerfüllung)**:
   - **Apology**: The student apologizes for being late ("entschuldigung, ich werde zu spät kommen").
   - **Explanation**: The student explains the reason for being late ("weil ich den Bus verpasst habe").
   - **Meeting Place**: The student mentions where Tobias should wait ("Treffen wir uns vor dem Park").

   The student has addressed all three points within the word limit.

   Score: 5
2. **Language (Sprache)**:
   - **Vocabulary and Structures**: The student uses appropriate vocabulary and structures, with minor errors ("Kannst du mir warten bitte?" should be "Kannst du auf mich warten bitte?").
   - **Coherence and Cohesion**: The message is coherent and logically structured. "Bis dann" should be "Bis dann" or "Bis später".

   Score: 3.5

Participant 2: Original text (Task part 1)



Participant 2: Transcribed text

Evaluation:

1. **Task Completion (Aufgabenerfüllung)**:
   - **Apology**: The student apologizes for being late ("es tut mir leid").
   - **Explanation**: The student explains the reason for being late ("weil ich im Stau bleibe").
   - **Meeting Place**: The student mentio[n]s a meeting place, though not clearly ("Du kannst mich ins Kino warten" should be "Du kannst im Kino auf mich warten").

   The student has addressed all three points within the word limit.

   Score: 5
2. **Language (Sprache)**:
   - **Vocabulary and Structures**: The student uses appropriate vocabulary but with grammatical errors ("weil ich im Stau bleibe" should be "weil ich im Stau stehe", "Du kannst mich ins Kino warten" should be "Du kannst im Kino auf mich warten", "Ich warte auf dein Antwort" should be "Ich warte auf deine Antwort").
   - **Coherence and Cohesion**: The message is generally coherent but has structural issues affecting clarity.

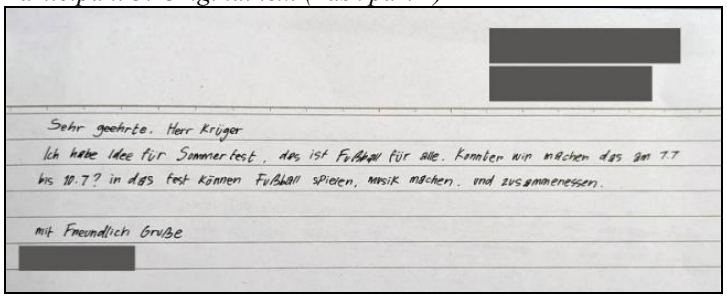
   Score: 2

Writing Task Part 2:

Herr Krüger, der Leiter von Ihrem Sportverein, möchte ein Sommerfest für Mitglieder organisieren und bittet Sie um Hilfe.
Schreiben Sie Herrn Krüger eine E-Mail:
- Schreiben Sie, wie Sie die Idee finden.
- Fragen Sie nach dem Datum.
- Schlagen Sie etwas vor, was man auf dem Fest machen kann.

Schreiben Sie 20-30 Wörter.
Schreiben Sie zu allen drei Punkten.

Participant 3: Original text (Task part 2)



Participant 3: Transcribed text

Student Text:

Sehr geehrter Herr Krüger,

Ich habe eine Idee für das Sommerfest: Fußball für alle. Können wir das vom 7.7 bis 10.7 machen? Auf dem Fest können wir Fußball spielen, Musik machen und zusammen essen.

Mit freundlichen Grüßen,
[redacted]

Participant 3: ChatGPT's Evaluation

Evaluation:

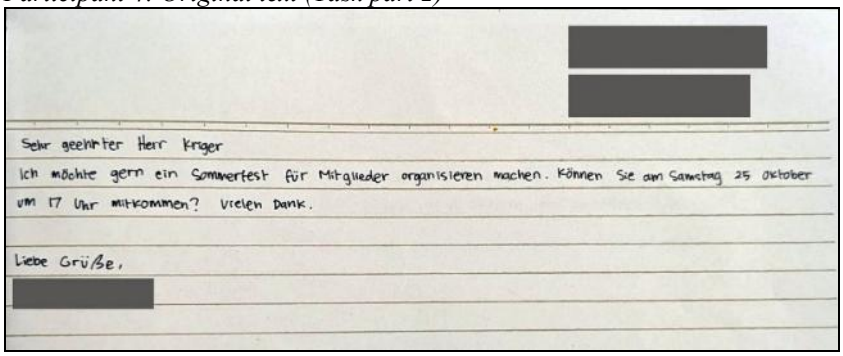
1. **Task Completion (Aufgabenerfüllung)**:
   - **Idea**: The student expresses their opinion about the idea ("Ich habe eine Idee für das Sommerfest: Fußball für alle").
   - **Date**: The student asks about the date ("Können wir das vom 7.7 bis 10.7 machen?").
   - **Suggestion**: The student suggests a[redacted]ities ("Auf dem Fest können wir Fußball spielen, Musik machen und zusammen essen").

   The student has addressed all three points within the word limit.

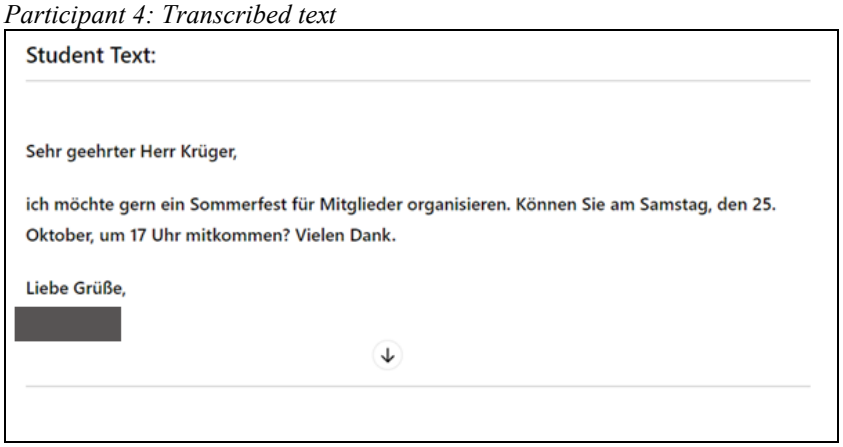
   Score: 5
2. **Language (Sprache)**:
   - **Vocabulary and Structures**: The student uses appropriate vocabulary but with several grammatical errors ("Sehr geehrte, Herr Krüger" should be "Sehr geehrter Herr Krüger", "Ich habe Idee" should be "Ich habe eine Idee", "Konnter wir machen" should be "Können wir das machen", "in das fest" should be "auf dem Fest", "zusammenessen" should be "zusammen essen").
   - **Coherence and Cohesion**: The message is generally coherent but has structural issues affecting clarity.

   Score: 2

Participant 4: Original text (Task part 2)



Participant 4: Transcribed text



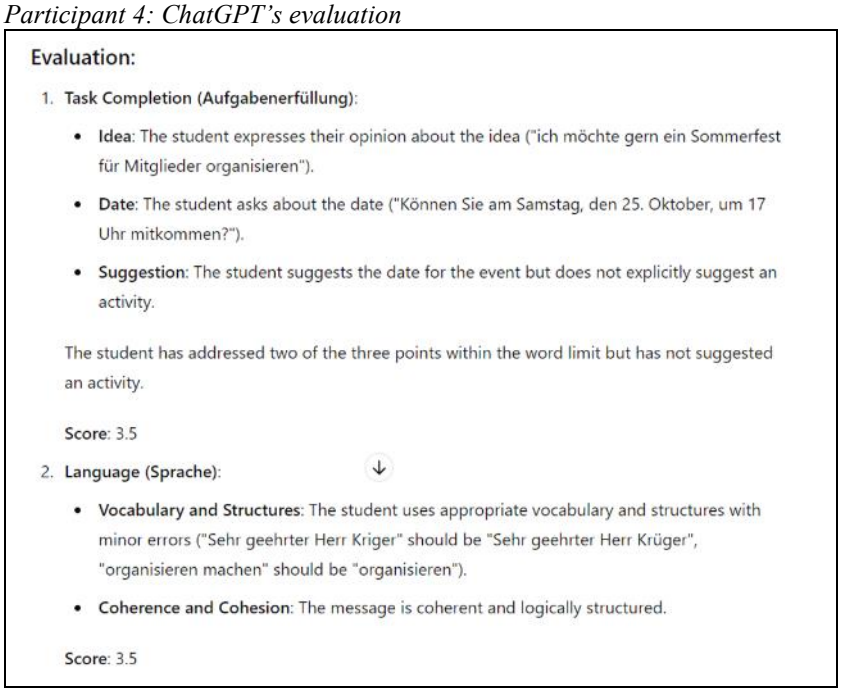
Student Text:

Sehr geehrter Herr Krüger,

ich möchte gern ein Sommerfest für Mitglieder organisieren. Können Sie am Samstag, den 25. Oktober, um 17 Uhr mitkommen? Vielen Dank.

Liebe Grüße,

Participant 4: ChatGPT's evaluation



Evaluation:

1. **Task Completion (Aufgabenerfüllung)**:
   - **Idea**: The student expresses their opinion about the idea ("ich möchte gern ein Sommerfest für Mitglieder organisieren").
   - **Date**: The student asks about the date ("Können Sie am Samstag, den 25. Oktober, um 17 Uhr mitkommen?").
   - **Suggestion**: The student suggests the date for the event but does not explicitly suggest an activity.

   The student has addressed two of the three points within the word limit but has not suggested an activity.

   Score: 3.5
2. **Language (Sprache)**:
   - **Vocabulary and Structures**: The student uses appropriate vocabulary and structures with minor errors ("Sehr geehrter Herr Kriger" should be "Sehr geehrter Herr Krüger", "organisieren machen" should be "organisieren").
   - **Coherence and Cohesion**: The message is coherent and logically structured.

   Score: 3.5