Received Date: 11 November 2025 I Accepted Date: 12 December 2025

RESEARCH ARTICLE

OPUS Journal of Society Research

Reliability of Human Expert and AI Raters in Translation Assessment

Yasemin Uzun[1]

[1] Assoc. Prof. Dr., Çanakkale Onsekiz Mart University, Faculty of Education. Çanakkale/Türkiye
ROR ID: https://ror.org/05rsv8p09
ORCID: 0000-0001-8995-772X
E-Mail: yaseminuzun@comu.edu.tr

November 2025
Volume:22
Issue:6
DOI: 10.26466/opusjsr.1821518

Citation:
Uzun, Y. (2025). Reliability of human expert and AI raters in translation assessment *OPUS– Journal of Society Research*, 22(6), 1305-1317.

Abstract

Although AI-based assessment systems offer new opportunities in education, their consistency with human judgment in measuring complex cognitive skills such as translation remains debatable. This study examines inter-rater reliability between a domain expert and AI raters (ChatGPT-5 and Gemini 1.5 Pro) in evaluating C2-level Turkish translations. Using a convergent mixed-methods design, translations from 14 students were scored with a 5-point analytic rubric. Krippendorff's alpha revealed low overall agreement (α = .392), particularly weak in "Semantic Accuracy" (α = .288). Qualitative analysis identified three key divergences: task fidelity, error severity perception, and criterion interpretation variability. Findings show AI models exhibit partial consistency in formal accuracy but systematically diverge from human experts in semantic nuance, style, and contextual appropriateness. The expert adopted a "task-oriented" approach, while AI models were more "form-focused" (Gemini) or "surface coherence-oriented" (ChatGPT). Although AI systems serve as useful auxiliary tools in translation assessment, they are not able to replace expert judgment

Keywords: *Artificial intelligence, inter-rater reliability, teaching Turkish as a foreign language, translation assessment.*

Öz

Yapay zekâ tabanlı değerlendirme sistemleri eğitimde yeni olanaklar sunsa da çeviri gibi karmaşık bilişsel becerilerin ölçümünde bu değerlendirme sistemlerinin insan yargısıyla tutarlılığı tartışmalıdır. Bu çalışma, C2 düzeyinde Türkçe çevirilerin değerlendirilmesinde alan uzmanı ile yapay zekâ puanlayıcıları (ChatGPT-5 ve Gemini 1.5 Pro) arasındaki puanlayıcılar arası güvenirliği incelemektedir. Yakınsak karma yöntem tasarımı kullanılarak, 14 öğrencinin çevirileri 5'li analitik rubrikle puanlanmıştır. Krippendorff alfa, düşük genel uyum (α = .392) ortaya koymuş, özellikle "Anlamsal Doğruluk" boyutunda uyum zayıf bulunmuştur (α = .288). Nitel analiz üç temel farklılık belirlemiştir: görev sadakati, hata ciddiyeti algısı ve kriter yorumlama çeşitliliği. Bulgular, yapay zekâ modellerinin biçimsel doğrulukta kısmi tutarlılık gösterdiğini ancak anlamsal nüans, üslup ve bağlamsal uygunlukta insan uzmanından sistematik olarak ayrıştığını ortaya koymaktadır. Uzman "görev odaklı" bir yaklaşım benimserken, yapay zekâ modelleri daha "biçim odaklı" (Gemini) veya "yüzeysel tutarlılık odaklı" (ChatGPT) değerlendirmeler yapmıştır. Yapay zekâ sistemleri çeviri değerlendirmesinde yararlı yardımcı araçlar olsa da uzman yargısının yerini alamamaktadır.

Anahtar Kelimeler: *Çeviri değerlendirmesi, yabancı dil olarak Türkçenin öğretimi, yapay zekâ, puanlayıcı güvenirliği.*

.

Review Note: *Evaluated by Double-Blind Peer Review*
Ethical Statement: *Ethical matters may be addressed via editorialoffice@opusjournal.net*

Introduction

Recent developments in artificial intelligence (AI) have fundamentally reshaped educational and assessment processes, offering novel alternatives for evaluating individual competencies. In recent years, AI-based assessments have generated substantial transformation in education and skill development. Such assessments are considered to provide high reliability in monitoring both interpersonal competencies and academic achievement and learning processes (Kotlyar & Krasman, 2025). Recent research demonstrates that large language models exhibit state-of-the-art performance in translation quality assessment (Kocmi & Federmann, 2023). AI systems offer opportunities to evaluate student performance more objectively and systematically through the application of big data analytics and machine learning techniques. By producing results comparable to those of human evaluators, AI systems enable educators to examine student performance in a more objective and structured manner (Kotlyar & Krasman, 2022).

AI systems provide significant advantages over traditional methods in assessment and feedback provision through characteristics such as objectivity, consistency, speed, and scalability. These qualities are particularly crucial in large-scale educational settings (Kotlyar & Krasman, 2025; Fahmy, 2024; Farrokhnia et al., 2024; Zawacki-Richter et al., 2019). Both traditional metric frameworks (Lommel et al., 2014) and neural network-based approaches (Rei et al., 2020) are employed in translation evaluation.

Moreover, the rapid feedback capabilities afforded by AI enhance the effectiveness of educational processes and support personalized learning experiences (Fahmy, 2024; Kotlyar & Krasman, 2025). The more effective utilization of AI-generated data by educators to monitor and guide student development may contribute to the improvement of learning processes. As the importance of interpersonal skills increases, AI feedback in the assessment process becomes increasingly significant. This situation underscores the necessity for more in-depth research on AI integration in education (Kotlyar & Krasman, 2025).

However, the proliferation of AI-based assessments has revealed certain disadvantages. Primarily, AI's capacity to comprehend emotional and cultural nuances remains limited. This inadequacy may result in assessments lacking the richness of human interaction. Furthermore, dependence on AI systems may lead teachers and students to overlook the human element in assessment processes. Human experience and intuition in evaluation processes are critically important in making sense of AI-generated data.

Within this framework, establishing a balance between the opportunities and limitations offered by AI is of paramount importance for understanding complex and dynamic structures such as the translation process. Translation is not merely the task of transferring words into another language. It is a multidimensional process requiring the effective communication of emotion, intention, and meaning (Bassnett, 2002). The translation process constitutes a complex form of communication emerging from the convergence of both linguistic and cultural elements, wherein the translator's role extends far beyond word transfer. Translators must endeavor to establish a cultural connection with readers, beyond merely conveying the meaning and significance of the foreign text to native readers (Venuti, 2012). In addition, the translator's role and textual analysis are significant in emphasizing social and cultural elements. Translation is simultaneously a meaning-making process within a socio-cultural context. Snell-Hornby (1988) articulates that the translation process functions as a bridge not only between languages but also between cultures. In this context, cultural transfers are required to be executed carefully to prevent losses or deformations in translation.

Rather than transferring every detail from the source text, the translator prioritizes the effect to be created upon the reader. Building on this approach, Munday (2016) emphasizes that more flexible and creative approaches have developed in translation practice. Reiss and Vermeer (1984) note that successful translation processes require consideration of the target audience and the purpose of translation. Due to the necessity of considering these two fundamental factors, functionality assumes prominence in translation.

When the effects of AI in education converge with the complexity of translation processes, despite certain inadequacies and disadvantages, new opportunities and approaches emerge for both educators and translators. In this context, this study will examine the Turkish translation competencies of international undergraduate students and compare AI and human assessments.

In contemporary contexts, foreign language learning holds critical importance in enabling individuals to enhance their global communication skills and engage with diverse cultures. "Teaching Turkish as a Foreign Language" is regarded as a strategic domain for imparting both language proficiency and cultural awareness (Özdemir, 2018). While the rich phonetic structure of Turkish facilitates linguistic encoding for foreign learners, it also enhances the language's significance at the international level (İşcan, 2011; Özdemir, 2018).

AI-based tools enable personalized and interactive learning experiences in foreign language teaching, rendering the process more effective by offering solutions tailored to learners' individual needs (Kaleli & Özdemir, 2025). Teaching Turkish as a foreign language not only imparts linguistic competence but also promotes cultural interaction. The rapid advancement of digital technologies has led to the displacement of traditional classroom-based methods by online platforms and AI-supported systems. AI's role in education possesses tremendous potential in evaluating student performance, providing personalized feedback, and delivering interactive learning experiences (Luo et al., 2025).

The purpose of this study is to determine the level of consistency and reliability between AI-based assessment systems and human raters (domain experts) in evaluating the Turkish translation competencies of international undergraduate students. Thus, AI systems' capacity to approximate expert judgment and methodological differences are comparatively revealed. Accordingly, the research seeks answers to the following questions:

a) What is the level of inter-rater reliability (IRR) between AI-based raters (ChatGPT and Gemini) and the domain expert?
b) At which points do the raters converge and diverge across the rubric dimensions of grammatical accuracy, semantic accuracy, fluency and naturalness, lexical choice, and cultural appropriateness?

Methodology

Research Design

This study was conducted according to a convergent mixed-methods design, examining the agreement between artificial intelligence (AI)-based raters (ChatGPT and Gemini) and a domain expert in evaluating C2-level student translation texts. The quantitative dimension of the research focused on inter-rater reliability (IRR) coefficients, while the qualitative dimension concentrated on comparative content analysis based on raters' "evidence" explanations included in the rubric.

Study Group and Data Collection Process

The study was conducted with 14 international undergraduate students proficient in Turkish at the C2 level, enrolled at a university in western Türkiye, who had completed course access/permission procedures. Convenience sampling was employed in participant selection, considering accessibility and ease of implementation (Büyüköztürk et al., 2020).

The sample group consists of individuals learning Turkish as a foreign language. Therefore, the text to be translated was selected from a more general topic that addresses the individual and social contributions of foreign language learning, rather than from a specialized field. The countries and native languages chosen by the participants for the translation text are as follows: 3 participants from Pakistan (Urdu), 1 from the Democratic Republic of the Congo (French), 1 from Kenya (English), 1 from Iran (Persian), 1 from Angola (Portuguese), 1 from Albania (Albanian), 1 from Afghanistan (Persian), 1 from North Macedonia (Albanian), 1 from Burundi (Kirundi), 1 from Somalia (Arabic), 1 from Madagascar (Malagasy), and 1 from Somalia (English).

The native languages of the 14 graduate students participating in the study belong to six different language families. Eight of the participants speak languages from the Indo-European language family (Urdu, Persian, French, English, Portuguese, and Albanian). Among these, three participants' native language is Urdu, two speak Persian, two speak English, two speak Albanian, one speaks French, and one speaks Portuguese. The remaining participants represent different language families. The target language, Turkish, belongs to the Ural-Altaic language family.

The data collection process was performed with approval decision number 21/114 obtained from the Social Sciences and Humanities Ethics Committee of a public university. During the data collection process, each student was provided with a text in their native language on the topic "Individual and Social Contributions of Foreign Language Learning." Students were asked to translate this text into Turkish. Analyses were conducted on the 14 Turkish texts obtained.

Assessment Tool and Raters

Assessment was performed using an analytical rubric developed by the researcher, comprising five sub-dimensions:

a) Grammatical Accuracy b) Semantic Accuracy c) Fluency and Naturalness d) Lexical Choice e) Cultural Appropriateness

Scoring was conducted on a scale from 1 (Very Poor) to 5 (Excellent).

Raters: (i) Domain Expert (experienced in Turkish language teaching), (ii) ChatGPT-5, (iii) Gemini-Pro 1.5.

AI Rater Settings: ChatGPT-5 and Gemini-Pro 1.5 were operated through the web interface by creating new accounts; both models were provided with the same rubric text and identical task instructions as system prompts. The models scored texts that had been stripped of file names containing clues about student identities/country information and had no access to other raters' scores/evidence.

Data Analysis

Quantitative analysis: Inter-rater agreement was calculated using Krippendorff's alpha (α) coefficient, appropriate for ordinal data. For statistical significance of results, 95% confidence intervals were considered; in interpretation, Krippendorff's (2004) threshold value of $\alpha < .667$ was accepted as "low reliability."

Qualitative analysis: Raters' explanations in the "Evidence" columns were examined through comparative content analysis to identify the causes of low agreement in quantitative findings. This analysis yielded themes such as task fidelity, error severity, and academic register.

Findings

Inter-Rater Reliability (IRR) Analysis Findings

Data were analyzed on an ordinal scale. As no missing observations were present, all data were included in the analysis. Results are presented in Table 1.

Table 1. Inter-Rater Reliability

Rubric Dimension	Krippendorff's α	%95 CI (Confidence Interval)
Grammatical Accuracy	$\alpha = .441$	[.268 - .600]
Semantic Accuracy	$\alpha = .288$	[.123 - .451]
Fluency and Naturalness	$\alpha = .428$	[.234 - .596]
Lexical Choice	$\alpha = .455$	[.281 - .609]
Cultural Appropriateness	$\alpha = .311$	[.102 - .498]

Examination of Table 1 reveals that inter-rater reliability was at a low level according to Krippendorff's (2004) standards. Krippendorff (2004) classifies $\alpha \geq .800$ as adequate reliability and $\alpha < .667$ as "low reliability" in social science research. Within this framework, the overall reliability coefficient of $\alpha = .392$ falls considerably below the recommended threshold. The highest level of agreement was observed in the Lexical Choice dimension ($\alpha = .455$), whereas the lowest agreement was found in the Semantic Accuracy dimension ($\alpha = .288$).

Comparative Score Analysis and Qualitative Findings by Rubric Dimensions

To visualize the causes underlying the low-level agreement patterns observed in the IRR analysis (see Table 1) and to identify points of rater divergence, score distributions and evidence-based interpretations are presented below for each rubric dimension. Qualitative analyses revealed that one of the primary sources of divergence among raters stemmed from differences in the perception of "Task Fidelity." In this study, task fidelity was operationally defined as the degree to which students fulfilled the translation task into the target language (Turkish) while preserving the academic register, terminology, and semantic integrity required by the source text, without resorting to summarization or rewriting.

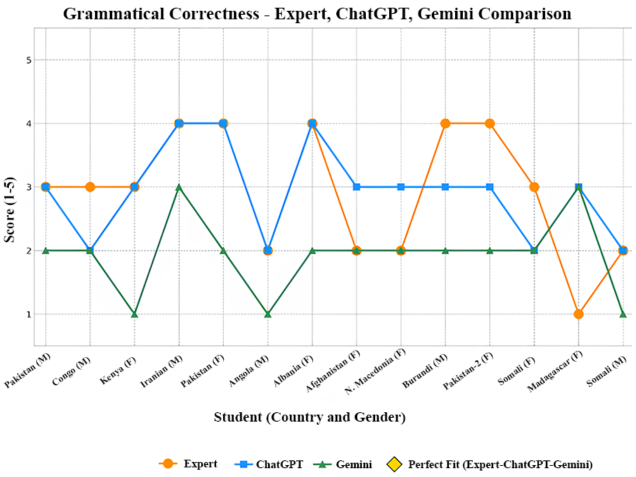

Figure 1. Comparison of Grammatical Accuracy

Figure 1 reveals the similarities and differences among the evaluations of three raters (Domain Expert, ChatGPT, Gemini) in the "Grammatical Accuracy" dimension. The findings generally demonstrate that Gemini (green line) exhibits a systematically lower scoring tendency compared to the other two raters (predominantly within the 1-2 point range). This situation indicates that Gemini adopts an approach that interprets error "severity" more stringently, prioritizing structural accuracy over surface-level errors.

Example 1 - Albanian (F) student: While the Domain Expert and ChatGPT evaluated this translation with a score of 4 (Good), Gemini assigned a score of 2 (Poor). The Domain Expert noted a limited morphological error with the statement "suffix usage is incorrect in some places," and ChatGPT reached a similar judgment with the comment "'Grammatical structure is strong, but suffix usage is occasionally erroneous.'" In contrast, Gemini classified expressions such as "gelişmesinde" as "serious phrase structure and verb voice errors," lowering the score by two points. This situation demonstrates that Gemini addresses grammar within a deeper syntactic framework.

Example 2 - Iranian (M) and Pakistani (M) students: In the Iranian example, both the Domain Expert and ChatGPT assigned a score of 4, identifying a translation that was semantically strong but contained minor formal errors. The Domain Expert reported the finding of "error in the use of the possessive suffix," while ChatGPT expressed the same error in a different dimension by stating "'Toplumsal bakımda' is not used in established Turkish." Similarly, for the Pakistani (M) student's text, both raters assigned a score of 3. While the Domain Expert noted "spelling errors are excessive: sampati kurmak, kabilyeti…," ChatGPT confirmed the same linguistic weakness using the rationale "suffix and syntax errors are frequently observed." These points of convergence demonstrate that the two raters exhibit high alignment in surface-level grammatical accuracy.

Example 3 - Madagascan (F) student (Methodological Divergence): The most pronounced methodological difference was observed in this text (Expert: 1, ChatGPT: 3, Gemini: 3). The Domain Expert indicated that the task was completely violated with the note "Text was not translated. Each paragraph was summarized," and assigned a score of 1 (Very Poor). In contrast, ChatGPT and Gemini disregarded this violation and scored only the surface-level grammatical accuracy of the existing text (e.g., ChatGPT - "'hafızayı günçlendirir' expression is erroneous, but the overall structure is correct"; Gemini - "The existing 3 sentences are fundamentally correct."). This situation demonstrates that both AI models did not include the task fidelity criterion in their evaluation.

In the Grammatical Accuracy dimension, the Domain Expert and ChatGPT detected errors similarly at the formal level, while Gemini conducted a more stringent evaluation by focusing on structural integrity; however, both AI models did not incorporate the text's appropriateness to the translation purpose (task fidelity) into their scoring. These results indicate that contextual awareness in measuring linguistic accuracy remains limited in AI-based assessments.

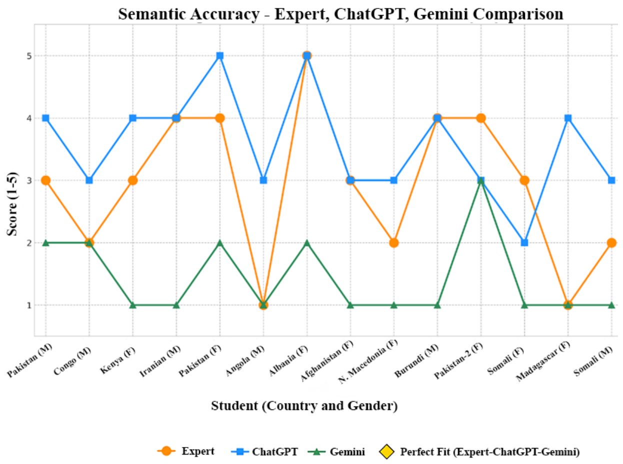

Figure 2. Comparative Analysis of Semantic Accuracy Scores

Figure 2 illustrates the evaluation patterns of three raters (Domain Expert, ChatGPT, Gemini) regarding meaning transfer in the "Semantic Accuracy" dimension. The findings demonstrate that this dimension reveals the highest methodological divergence among the three raters. While the Domain Expert and ChatGPT exhibited partial convergence by prioritizing the general flow of meaning and main idea coherence in texts, Gemini adopted a more stringent and rule-based evaluation model that foregrounded principles of complete fidelity to the source text and conceptual integrity.

Example 1 - Madagascan (F) student (Task Fidelity Divergence): This text represents the example where the most pronounced inter-rater conflict occurred. The Domain Expert and Gemini determined that the student had not fulfilled the translation task, and both assigned a score of 1 (Very Poor). The Domain Expert's note stated, "Text was not translated. Each paragraph was summarized"; Gemini's evidence was "This is not a translation; 95% of the academic content in the source text is lost." In contrast, ChatGPT disregarded this task fidelity violation and rewarded the student's summarization success, assigning a score of 4 (Good) ("Three paragraphs correctly summarize the main ideas…"). This situation demonstrates that AI models interpret task definition differently; ChatGPT prioritizes functionality, while Gemini emphasizes formal fidelity.

Example 2 - Albanian (F) student: While the Domain Expert and ChatGPT evaluated this translation with a score of 5 (Excellent), Gemini assigned only 2 (Poor) points. The Domain Expert commented "Meaning is very well conveyed"; ChatGPT similarly assessed "The text holistically conveys accurate meaning"; however, Gemini assigned a low score with the explanation "The text is more of an incomplete summary than a translation. Key concepts such as 'cognitive flexibility', 'problem solving', 'analysis', and 'multitasking' have been omitted." This comparison demonstrates that Gemini prioritizes conceptual integrity, while the other two raters prioritize communicative adequacy.

Example 3 - Somali (F) student (Critical Error Assessment): In this text, the translation of "In a world where Bitcoin usage is increasing" instead of "In a globalizing world" was identified as an error by all three raters; however, the severity of the error was assessed differently. Gemini evaluated this situation as a critical error, stating "meaning and context are completely lost," and assigned 1 point. ChatGPT assigned 2 points, while the Domain Expert assigned 3 points, stating "Adequate but context is weak." This situation demonstrates that Gemini adopts a zero-tolerance approach to critical errors, while the other two raters adopt a flexible evaluation style that considers holistic meaning preservation.

Example 4 - Afghan (F) student: The Domain Expert and ChatGPT assigned 3 points due to the excessive number of "(………)" gaps in the text; however, Gemini interpreted the reversal of the source text's expression "is supported by scientific findings" to "This subject does not encompass scientific findings" as a semantic contradiction and evaluated it with 1 point. This example indicates that Gemini classifies contextual contradictions as critical errors and establishes a stricter tolerance threshold for semantic deviations.

The semantic accuracy dimension demonstrates that the criteria used by AI models in meaning transfer are fundamentally different from each other. While ChatGPT and the Domain Expert employ a more holistic scoring approach based on the preservation of the main idea and overall semantic coherence of the text, Gemini exhibits a more systematic yet stringent model in terms of source text fidelity, conceptual completeness, and critical error sensitivity. This difference indicates that the definition of "semantic accuracy" in AI-based translation assessments has not yet been standardized and cannot fully reflect the cognitive flexibility of human raters.

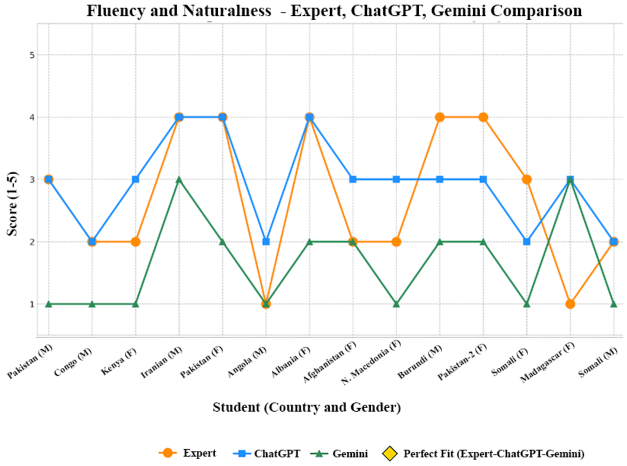

Figure 3. Fluency and Naturalness Dimension Scoring

Figure 3 illustrates the patterns among the evaluations of three raters (Domain Expert, ChatGPT, Gemini) regarding the "Fluency and Naturalness" dimension. The findings reveal two distinct tendencies in this dimension:

(1) High-level perceptual convergence between the Domain Expert and ChatGPT regarding the natural flow of the text,

(2) Gemini's adoption of a more stringent evaluation approach that methodologically diverges from the other two raters and associates "fluency" not only with formal smoothness but also with the reflection of the source text's academic register.

Example 1 - Pakistani (M) student: The Domain Expert and ChatGPT assigned this text a score of 3 (Moderate). Both raters provided similar justifications by referring to the same unnatural expressions:

Domain Expert Evidence: "Sentences are weak in terms of naturalness and fluency characteristics: kafama (akıllı) daha geliştirmek içindir…"

ChatGPT Evidence: "Some sentences are far from natural flow and sound artificial in Turkish. Expressions like 'sampati kurmak, kafama (akıllı)…' distort meaning."

These points of convergence demonstrate that the two raters evaluate the concept of "naturalness" through similar linguistic indicators and employ a common reference framework in their perception of fluency. Similar convergence was additionally observed in the (4-Good) scores assigned to texts by Iranian (M) and Pakistani (F) students.

Example 2 - Gemini's systematic divergence: Gemini's (green line) scores are generally lower. The reason for this difference is that Gemini associates "fluency" not only with the surface-level smoothness of sentences but also with the preservation of the original text's academic tone.

For instance, Gemini assigned 3 points to the Iranian (M) text that the Domain Expert and ChatGPT evaluated as "generally fluent" (Score 4).

Gemini Evidence: "Existing sentences are short, simple, and comprehensible. However, the entire text remains too 'simplistic'; the fluent and academic language of the original text has not been reflected."

Similarly, having identified the phenomenon of "word-for-word translation" in the Pakistani (M) student's text ("Sentences are very mechanical and give the impression of word-for-word translation."), Gemini evaluated this situation as a critical error disrupting fluency and assigned a score of 1 (Very Poor).

Example 3 - Madagascan (F) student (Task Fidelity Divergence): This text presents a notable example of methodological transition among raters. Having agreed with the Domain Expert (Score 1) in the "Semantic Accuracy" dimension, Gemini this time assigned a score similar to ChatGPT (Score 3) in the "Fluency" dimension.

Domain Expert Evidence: "Text was not translated; only a summary was written."

ChatGPT Evidence: "Sentences are very simple, but fluent within themselves."

Gemini Evidence: "Although the text is of a summary nature, sentence structure is clear."

This example demonstrates that both AI models completely disregarded the principle of task fidelity and focused solely on the formal flow of the text.

The "Fluency and Naturalness" dimension serves as an important indicator revealing the extent to which human and AI-based evaluations overlap. While the Domain Expert and ChatGPT evaluate fluency along the axes of linguistic naturalness and readability, Gemini associates this dimension with the transfer of academic tone, structural complexity, and formality in lexical choice. Consequently, the manner in which AI models measure fluency operates independently of task context (translation/summary distinction), indicating that they are not able to adequately account for the contextual dimensions of linguistic performance.

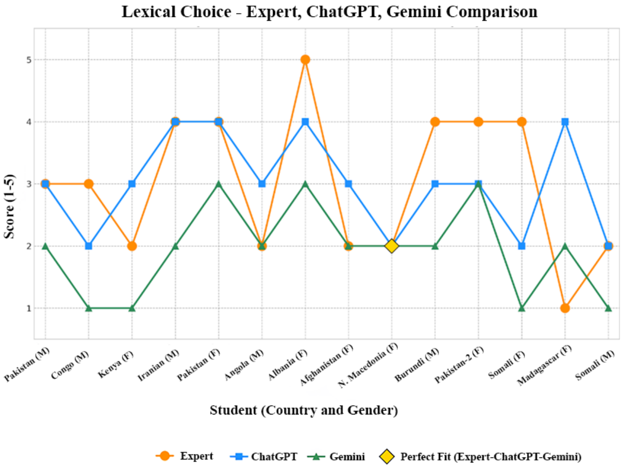

Figure 4. Lexical Choice Dimension Scoring

Figure 4 demonstrates that inter-rater evaluations in the "Lexical Choice" dimension simultaneously exhibit examples of both perfect agreement and methodological conflict. This dimension contains the only instance among 14 students and 5 sub-dimensions (totaling 70 evaluation points) where all three raters assigned the same score ("North Macedonia (F)"); conversely, it reveals substantial philosophical divergences in the "Somali (F)" and "Madagascan (F)" texts.

Example 1 - North Macedonian (F) student (Perfect Agreement Example): All three raters evaluated this text with a score of 2 (Poor). However, they reached the same conclusion through different justifications:

Domain Expert: Emphasized that academic concepts in the source text were conveyed with incorrect words through the expression "'Bir dil öğrenerek daha akılı bir kişi olabilirsin...'"

ChatGPT: Focused on formal and orthographic errors, stating "Words like 'Biyininizi', 'profesiorel' do not exist in Turkish."

Gemini: Identified terminological errors with examples such as "'Fen bilimler' (instead of Eğitim bilimleri), 'Diğer anda' (instead of Öte yandan)."

This situation demonstrates that raters may converge on the same conclusion even when following different diagnostic pathways, and that weak lexical choice might be consistently identified at the perceptual level.

Example 2 - Somali (F) student (Critical Error Divergence): This text illustrates how differently raters perceive "critical error" regarding lexical choice. Scores were distributed across a wide range: 4 (Expert), 2 (ChatGPT), and 1 (Gemini).

Gemini: Evaluated this error as fatal, stating "For an academic text, choosing the word 'Bitcoin' is the greatest error, demonstrating that the fundamental context was not understood at all" (Score: 1).

ChatGPT: Associated the error with contextual deficiency through the statement "Mixed English-Turkish usage such as 'Dünya pazarlama / ekonomi…' is present," and assigned 2 points.

Domain Expert: Despite the observation "Instead of 'Küreselleşen dünyada', the expression 'Bitcoin...' was used," assigned 4 points. However, this scoring exhibits serious internal inconsistency with the "not adequate" statement in their own evidence column.

Example 3 - Madagascan (F) student (Task Fidelity Divergence): This text demonstrates how evaluation of lexical choice independent of task context produces different results among raters.

Domain Expert (Score 1): Invalidated the text for not fulfilling the translation task, stating "Text was not translated."

ChatGPT (Score 4): Disregarded task fidelity and evaluated only the word choice within the summary text itself, noting "Lexical choice is generally appropriate."

Gemini (Score 2): Unlike ChatGPT, compared the summary text's words with the source text's academic terminology and assigned 2 points with the justification "None of the source text's rich academic terminology has been utilized."

This situation demonstrates that AI models employ different reference frameworks even when evaluating a "summary" text: ChatGPT prioritizes internal consistency, while Gemini bases evaluation on comparative terminological adequacy with the source.

The "Lexical Choice" dimension exhibits a hybrid pattern where the three raters employ different cognitive strategies for evaluation yet can reach the same conclusion in some cases. ChatGPT's more text-internal consistency-focused evaluations and Gemini's source text fidelity and academic terminology-focused assessments demonstrate that AI models prioritize different criteria compared to human experts in the lexical choice dimension. These findings reveal that the "lexical choice" criterion, particularly in translation-based assessment applications, needs to be redefined according to contextual meaning and task fidelity variables.

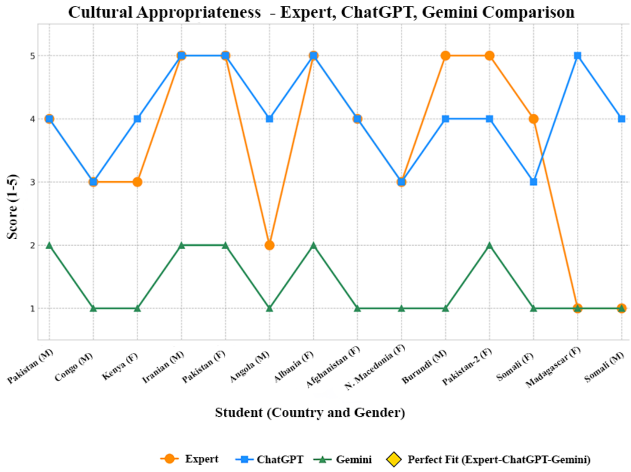

Figure 5. Comparative Analysis of Cultural Appropriateness Scores

Figure 5 illustrates the most profound methodological divergence among raters in the dimension of "Cultural Appropriateness." Unlike the previous four dimensions, this dimension is grounded not only in the level of error detection but also in a philosophical interpretation difference regarding the definition of the rubric items. Upon examining the "Evidence" columns of the raters, it becomes evident that each rater interprets the criterion of "Cultural Appropriateness" in three different manners: thematic/universality, academic style conveyance, and task fidelity.

Thematic Assessment (Domain expert and ChatGPT)

The Domain expert (orange) and ChatGPT (blue) lines exhibit a high correlation in the range of 3-5 points. Both raters interpret the concept of "cultural appropriateness" through the lens of the significance of the translated text for Turkish readers and adherence to universal themes.

ChatGPT Evidence (Iran E, Score 5): "Cultural transmission is very close to Turkish; the meaning is compatible on a universal level."

Domain expert Evidence (Pakistan K, Score 5): "Cultural context is appropriate."

These examples indicate that both raters associate cultural appropriateness at a content level, specifically relating to universal themes such as "the benefits of language learning" and "intercultural interaction."

Stylistic Assessment (Gemini)

Gemini (green line), on the other hand, has consistently assigned scores that are 1-2 points lower for the same texts. The reason for this lies in Gemini's definition of the concept of "culture" as the conveyance of academic culture. Thus, cultural appropriateness is sought not in the content of the text but at the formal and discursive levels (formality, terminology, academic tone).

Gemini Evidence (Pakistan E, Score 2): "The culture of the source text is academic and formal. The translation fails to maintain this formality, completely distancing itself from the context with the use of informal language such as 'head.'"

This difference in approach arises from Gemini's view that academic tone and formal consistency are integral parts of cultural transmission. Consequently, according to Gemini, a thematically suitable but informally structured text is not considered culturally "appropriate."

Divergence in Task Fidelity

The most pronounced conflict in the dimension of cultural appropriateness has been observed in texts with violations of task fidelity. The examples of "Madagascar (K)" and "Somali-2 (E)" notably represent this conflict.

Example - Madagascar (K) Student: This text shows the most extreme score disparity among the 210 evaluation points (Expert: 1, ChatGPT: 5, Gemini: 1).

Domain expert (Score 1): "The text has not been translated," deeming the evaluation invalid due to a violation of the task.

Gemini (Score 1): "The academic culture of the text has been completely destroyed; the formal tone has been lost with an excessively simplified summary."

ChatGPT (Score 5): Ignoring the task violation, it awarded the thematic coherence of the summary, stating that "There are no issues in the Turkish context; the listing of benefits is understandable."

Example - Somali-2 (E) Student: A similar pattern can be observed in this text as well (Expert: 1, ChatGPT: 4, Gemini: 1).

Domain expert : "Insufficient."

Gemini: "The formal, objective, and academic tone has been entirely eradicated."

ChatGPT: "The emphasis on culture and empathy is suitable for the Turkish reader."

This table illustrates that ChatGPT prioritizes thematic coherence, while Gemini and the Domain expert prioritize adherence to task definition and stylistic integrity.

The dimension of "Cultural Appropriateness" has revealed that the criterion definitions in the rubric are understood differently by the three raters across three distinct levels:

Domain expert and ChatGPT: Evaluated cultural appropriateness from thematic and communicative perspectives.

Gemini: Interpreted culture through formal and academic style dimensions.

These results indicate that abstract criteria such as cultural appropriateness are susceptible to issues of consistency in AI-supported evaluations. AI models struggle to differentiate the "semantic," "formal," and "contextual" dimensions of culture, thereby failing to fully reflect the intuitive and sociocultural sensitivity of human experts.

Discussion and Conclusion

In this study, no statistically reliable alignment was found between AI-based raters (ChatGPT and Gemini) and the domain expert in the evaluation of translation texts from C2-level students (Overall α = .392). In this context, the failure to exceed the $\alpha \geq .667$ threshold, which is accepted in social sciences (Krippendorff, 2004), indicates that inter-rater consistency is at a "low" level. This finding suggests that AI models do not possess the level of reliability required to replace human expert judgment in scoring. However, beyond the quantitative findings, qualitative analysis indicates that this low alignment is not coincidental; rather, it stems from three distinct systematic divergences, which are explained below.

Criterion Interpretation Divergence: "Academic Style" and "Thematic Universality"

The lowest reliability values among the rubric dimensions were recorded for "Cultural Appropriateness" (α = .311) and "Semantic Accuracy" (α = .288). This situation aligns with the difficulties AI models encounter when evaluating abstract and interpretation-dependent criteria (Tang et al., 2024). This study demonstrates that AI systems are capable of provide relatively consistent scores at the linguistic idea and structure level in written products; however, they fail to reflect the intuitive judgment of human raters regarding contextual, cultural, and syntactic nuances. In this regard, the domain expert and ChatGPT evaluated the text in terms of meaningful transmission to Turkish and thematic universality, while Gemini interpreted the same criterion in terms of formally conveying the academic style of the source text; this led to differing evaluations of the same metric by two raters.

Error Severity Divergence: "Form-Focused" vs. "Communicative Competence"

The analyses revealed distinct "rater profiles" in the scoring behaviors of the AI models. For instance, Gemini adopted an approach centered on formal accuracy and adherence to rules, showing lower tolerance for meaning errors. This orientation has been supported by a tendency to assign low scores for contextually critical errors. In contrast, ChatGPT prioritized the overall flow and coherence of the text, following a more communicative evaluation line. This difference points to basic issues regarding validity and reliability in AI-based automated or semi-automated scoring systems (Doewes & Pechenizkiy, 2021).

Methodological Divergence: Task Fidelity vs. Superficial Quality?

The most profound divergence among raters emerged in instances where students paraphrased instead of translating. While the domain expert and Gemini categorized this as a task violation, scoring it low, ChatGPT described the same example as a "superficially successful summary," assigning a higher score. This situation demonstrates that AI models focus solely on the surface of the text without adequately considering the analytical structures of the rubric. From this perspective, it cannot be claimed that AI models fully emulate human intuition and evaluation context (Uyar & Büyükahıska, 2025).

In conclusion, AI models have the ability of producing relatively consistent results in superficial linguistic accuracy and fluency but face reliability issues in deeper cognitive dimensions, such as meaning, style, and cultural appropriateness. The evaluations from the domain expert contained a more holistic "task-oriented" judgment, while the AI models remained focused on formal accuracy (Gemini) or internal consistency (ChatGPT). Therefore, it concludes that AI-based evaluation systems may be used as supportive raters at this stage but are unable to replace human experts in final judgments. This result suggests that future studies should focus on directed training (prompt calibration) and hybrid human-AI scoring systems to enhance the sensitivity of AI models to rubric criteria. Furthermore, applications conducted at different language levels (A2-C2) and across various task types are anticipated to comprehensively test the scoring stability of AI.

Limitations

Sample Size and Generalizability: The study's findings are based on a specific group of C2-level students and their translation texts. A limited sample size may hinder the generalizability of the results to broader populations or different educational contexts.

Rater Diversity: The study primarily focused on two AI-based raters (ChatGPT and Gemini) and one human expert. The lack of diversity in raters may limit the scope of the findings, as incorporating additional human raters with varied backgrounds or expertise could yield different results.

Contextual Factors: The evaluation of translation texts can be influenced by various contextual factors, including the specific content of the texts, the cultural background of the students, and the evaluative criteria used. These factors may not have been fully accounted for in the study, which could impact the reliability of the findings.

Declarations

Funding: No funding was received for conducting this study.

***Conflicts of Interest**:* The author declares no conflict of interest.

***Ethical Approval**:* The study was conducted with the approval decision numbered 21/114 dated 24.10.2025 of the Social and Human Sciences Ethics Committee of a state university

***Informed Consent**:* Informed consent was obtained from all subjects involved in the study..

***Data Availability**:* The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request

***AI Disclosure**:* No artificial intelligence-based tools or applications were used in the conception, analysis, writing, or preparation of figures for this study. All content was generated by the author in accordance with scientific research methods and academic ethical standards.

References

Bassnett, S. (2002). *Translation studies*. Routledge.

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2020). *Bilimsel araştırma yöntemleri* (27. bs.). Pegem Akademi.

Doewes, A., & Pechenizkiy, M. (2021). On the limitations of human-computer agreement in automated essay scoring. *Proceedings of the 2021 Educational Data Mining Conference*.

Fahmy, Y. (2024). *Student perception on AI-driven assessment: Motivation, engagement and feedback capabilities* [Yüksek lisans tezi, University of Twente]. University of Twente Student Theses. https://essay.utwente.nl/91297/

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, *61*(3), 460-474. https://doi.org/10.1080/14703297.2023.2195846

İşcan, A. (2011). Türkçenin yabancı dil olarak önemi. *International Journal of Eurasia Social Sciences*, 2(4), 29-36.

Kaleli, S., & Özdemir, A. (2025). Artificial intelligence and its role in teaching Turkish as a foreign language. *Turkish Linguistics Journal*.

Kocmi, T., & Federmann, C. (2023). Large language models are state-of-the-art evaluators of translation quality. Proceedings of the 24th Annual Conference of the European Association for Machine Translation, 193-203. https://aclanthology.org/2023.eamt-1.19/

Kotlyar, I., & Krasman, J. (2022). Virtual simulation: New method for assessing teamwork skills. *International Journal of Selection and Assessment,* *30*(3), 344-360. https://doi.org/10.1111/ijsa.12368

Kotlyar, I., & Krasman, J. (2025). Student reactions to AI versus human feedback in teamwork skills assessment. *International Journal of Educational Technology in Higher Education, 22*(1), 1-34. https://doi.org/10.1186/s41239-025-00555-9

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2. bs.). Sage Publications.

Lommel, A., Burchardt, A., & Uszkoreit, H. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. Tradumàtica: Tecnologies de la Traducció, 12, 455-463. https://doi.org/10.5565/rev/tradumatica.77

Luo, J., Zheng, C., Yin, J., & Teo, H. H. (2025). Design and assessment of AI-based learning tools in higher education: A systematic review. International Journal of Educational Technology in Higher Education, 22, 42. https://doi.org/10.1186/s41239-025-00540-2

Munday, J. (2016). *Introducing translation studies: Theories and applications*. Routledge.

Özdemir, C. (2018). Günümüzde yabancı dil olarak Türkçe öğretiminin durumu. *Alatoo Academic Studies, 18*(1), 11-19.

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2685-2702. https://doi.org/10.18653/v1/2020.emnlp-main.213

Reiss, K., & Vermeer, H. J. (1984). *Grundlehren einer allgemeinen Translationstheorie*. Cornelsen.

Snell-Hornby, M. (1988). *Translation studies: An interdisciplinary approach*. John Benjamins.

Tang, X., Chen, H., & Lin, D. (2024). Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Frontiers in Education, 9*, Article 11305227. https://doi.org/10.3389/feduc.2024.11305227

Uyar, A. C., & Büyükahıska, D. (2025). Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT. *International Journal of Assessment Tools in Education, 12*(1), 20-32. https://doi.org/10.21449/ijate.1517994

Venuti, L. (Ed.). (2012). *The translation studies reader* (3. bs.). Routledge.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education-where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 1-27. https://doi.org/10.1186/s41239-019-0171-0

Appendix A

Rubric

1 = Very Poor | 2 = Poor | 3 = Fair | 4 = Good | 5 = Excellent

Dimension	**Definition**	**Score (1–5)**	**Explanation / Evidence (example from the text, short note)**
Grammatical Accuracy	Compliance of the translation with Turkish grammar rules (suffixes, tense, syntax).		
Semantic Accuracy	The extent to which the meaning of the source text is accurately and completely conveyed.		
Fluency and Naturalness	The degree to which sentences are constructed naturally and read smoothly in Turkish.		
Word Choice	Appropriateness of vocabulary and terminology in context.		
Cultural Appropriateness	The extent to which cultural nuances are properly conveyed in Turkish.		

Appendix B

Text to be Translated

The Individual and Societal Contributions of Foreign Language Learning

In a globalizing world, knowing a foreign language plays a critical role in both individuals' personal development and professional success. Learning a foreign language is not only about acquiring an additional means of communication but also about understanding different cultures, developing empathy, and enhancing cognitive flexibility. Language learning improves an individual's thinking skills, strengthening their abilities in problem-solving, analysis, and multitasking.

Research within the field of educational sciences has shown that foreign language learning has positive effects on brain development. In particular, language education begun at an early age positively influences native language proficiency and enhances overall linguistic awareness. Moreover, studies indicate that language learning in adulthood can delay cognitive aging and help preserve memory capacity.

At the societal level, individuals who know foreign languages serve as bridges in intercultural interactions, fostering social cohesion and facilitating international collaboration. Economically, multilingual individuals gain a competitive advantage in the global market and have greater employment opportunities across different countries.