Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Assessing additional language oral production with ChatGPT: Validity, differences, and agreement with human raters

Carolina Poussier Milhau

cpoussier@uda.ad

University of Andorra https://orcid.org/0009-0002-9456-8512

Virginia Larraz Rada

University of Andorra https://orcid.org/0000-0002-3983-2117

Ferran Adell Español

Open University of Catalonia https://orcid.org/0000-0002-4192-3325

Nadia Azzouz Boudadi

University of Andorra https://orcid.org/0000-0002-8799-1681

Research Article

Keywords: Generative artificial intelligence, ChatGPT, Automated language assessment, Oral production, Assessment validity, Additional language acquisition, Human-AI comparison

Posted Date: May 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9768480/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

The integration of generative artificial intelligence into educational assessment raises questions about its validity, particularly for oral production, which has received limited attention compared to written tasks. This study investigates the validity of ChatGPT as a tool for assessing oral production in additional language acquisition within a higher education context. Twenty-six French A2 oral productions were analysed. Each production was assessed by two human raters based on audio recordings and by ChatGPT using verbatim transcripts, applying shared analytic criteria across four dimensions: content and development, coherence and organisation, syntactic control, and grammatical accuracy. Statistical analyses included the Shapiro–Wilk test, Wilcoxon signed-rank test, Spearman correlation, and Bland–Altman analysis. The results show a statistically significant difference between artificial intelligence-generated and human scores ($V = 36$, $p < .001$), with ChatGPT assigning lower scores on average (mean difference = − 1.02). The distribution of differences did not significantly deviate from normality ($W = .966$, $p = .519$). A very strong positive correlation ($\rho = .95$, $p < .001$) indicates strong agreement in the relative ranking of learners. These findings suggest that ChatGPT can reproduce consistent patterns in learner performance when assessment is restricted to text-accessible criteria, while remaining limited by the absence of prosodic and interactional information. The study supports the integration of generative artificial intelligence within hybrid assessment models combining automated analysis and human judgement.

Introduction

The integration of generative artificial intelligence (GenAI) into higher education has raised new questions about teaching, learning, and assessment practices. Recent developments in large language models (LLMs), such as ChatGPT, have expanded the possibilities for analysing student production and generating responses that resemble human judgement, prompting growing interest in their use for educational assessment (Xia et al., 2024).

Assessment in higher education has traditionally relied on human judgement, especially in tasks requiring interpretative assessment, such as open-ended responses or oral production. While human assessment allows contextualised and flexible judgement, it is also time-consuming and subject to variability between raters (Jukiewicz & Wyrwa, 2026). In this context, GenAI has been examined both as a challenge for academic integrity and as a potential tool to support assessment processes, particularly when assessment is guided by explicit rubrics and well-defined criteria (Abdullah et al., 2025; García-Varela et al., 2025; Luo et al., 2025; Moorhouse et al., 2023).

Despite this growing body of research, the use of GenAI for assessing oral production remains comparatively underexplored. This gap is relevant because oral competence involves linguistic, prosodic, and interactional dimensions that are not equally accessible to automated systems. Current AI-based approaches generally rely on the transcription of speech into text, which transforms the original oral data

and restricts assessment to textually observable features (OpenAI, 2024; Wolf et al., 2023). As a result, key aspects of oral performance, such as fluency, pronunciation, and prosody, may not be fully captured.

Given these limitations, empirical research is needed to examine how AI-generated scores compare with human assessment in oral assessment contexts. More specifically, it remains unclear whether differences between AI and human scoring are systematic, whether AI systems preserve the relative ranking of learners, and to what extent transcript-based AI assessment can be considered valid for evaluating oral production.

The present study examines the validity of ChatGPT as a tool for assessing oral production in additional language learning within a higher education context. It compares scores generated by ChatGPT with those assigned by human raters on French speaking tasks, focusing on score differences, agreement in learner ranking, and the extent to which ChatGPT can function as a reliable support tool in oral assessment.

State of the Art

1. GenAI and the transformation of assessment in higher education

Recent literature has increasingly focused on the impact of GenAI on higher education, particularly regarding assessment practices and the evolving role of university teachers in this context (Chan & Colloton, 2024; Luo et al., 2025). Assessment in university settings has historically relied on manual procedures that require a significant investment of time and are subject to variation in judgement (Xia et al., 2024). These challenges become particularly evident in large cohorts, where the volume of student production increases the workload associated with grading and may affect the consistency of assessment processes (Luo et al., 2025; Xia et al., 2024). From a theoretical perspective, assessment can be understood as an evidentiary reasoning process linking observed performances to inferences about underlying competences, as conceptualised in evidence-centered design (ECD) (Mislevy et al., 2003). In this sense, the introduction of GenAI does not simply affect grading procedures, but modifies the way evidence is generated, represented, and interpreted within the assessment process.

In this context, the emergence of GenAI has prompted institutions to reconsider existing assessment models. It has been demonstrated by numerous studies that the potential for students to utilise GenAI tools to generate responses or complete academic tasks gives rise to concerns regarding the validity and reliability of specific assessment formats (Audran, 2024; Carton & Ghyselen, 2024; Chan & Colloton, 2024; Luo et al., 2025; Moorhouse et al., 2023). This scenario underscores a broader pedagogical tension between the necessity to cultivate AI-related competencies and the risk that unregulated or uncritical utilisation of such tools may impede the development of authentic knowledge and skills, particularly in contexts where issues of academic integrity are at stake (Chan & Colloton, 2024). At the same time,

these developments reveal structural limitations in assessment models based exclusively on human judgement. Recent research suggests that GenAI contributes to a reconfiguration of assessment practices by intervening in different components of the evidentiary chain (Carton & Ghyselen, 2024). Notably, AI-based systems operate primarily on observable data and apply scoring procedures based on explicitly defined criteria, which makes their performance highly dependent on task design and data representation (Carton & Ghyselen, 2024; Chan & Colloton, 2024; Hao, 2025; Xia et al., 2024). From this perspective, the integration of GenAI requires reconsidering not only assessment formats, but also the alignment between tasks, evidence, and assessment criteria that underpin valid assessment.

This paradigm shift is evidenced by a growing body of research investigating the potential of GenAI to support instructors. The extant literature suggests that, when guided by explicit and structured criteria, GenAI may contribute to reducing grading workload and to improving the consistency of assessment (Ferran, 2025; Gaballo, 2024). Nevertheless, this research domain remains underdeveloped in comparison to studies that concentrate on the regulation of student utilisation of GenAI (Audran, 2024; Carton & Ghyselen, 2024; Moorhouse et al., 2023). These developments are particularly relevant in the assessment of language skills, where assessment requires detailed analysis of individual productions. In this domain, oral production poses specific challenges, as it involves time-intensive processes that are difficult to sustain in large cohorts (Jukiewicz & Wyrwa, 2026). Moreover, when oral performances are mediated through transcription, the available evidence is reduced to text-based representations, which further illustrates how the integration of GenAI transforms the nature of the evidence on which assessment decisions are based (Hao, 2025). Consequently, there is increasing interest in AI-assisted systems that can support the assessment of spoken language, while raising new questions about the validity and scope of such approaches.

2. Automated assessment with language models

Automated assessment based on LLMs has emerged as a growing field of research in higher education. Recent studies have investigated the use of LLMs, such as ChatGPT, to analyse student productions and generate scores based on predefined criteria (Abdullah et al., 2025; Flodén, 2025; Jukiewicz, 2024; Jukiewicz & Wyrwa, 2026). These systems introduce new methodological possibilities for educational assessment. From an assessment perspective, such developments can be understood as a reconfiguration of the evidentiary processes through which student performance is interpreted, as automated systems rely on explicitly defined criteria and observable data to generate scoring decisions (Hao, 2025).

2.1 Language models and open-ended assessment

Within the domain of educational assessment, a common distinction is made between closed-ended and open-ended tasks. While closed-ended tasks allow objective scoring procedures, open-ended tasks require interpretative judgement and are therefore more complex to assess, leading to increased

variability and reduced scoring stability, particularly in AI-based assessment contexts (García-Varela et al., 2025; Jukiewicz, 2024; Xia et al., 2024; Zubaer et al., 2025). This variability is closely linked to the interpretative nature of such tasks, as raters must rely on partially subjective criteria when making qualitative judgements (Alderson et al., 1995). In response to these challenges, recent research has explored the use of GenAI as a support tool for evaluating open-ended responses. For instance, García-Varela et al. (2025) demonstrate that ChatGPT can be utilised to assess written work and offer feedback, implying that AI could assist in overcoming the time-consuming and inconsistent nature of human assessment. However, they also state that, without a detailed, categorised framework, ChatGPT produces inconsistent results and criteria that vary from one interaction to the next. This reinforces the argument regarding dependence on the design of the assessment.

The effectiveness of these systems, however, depends on the explicit definition of the assessment framework. The literature has a recurring tendency to emphasise that AI-generated scoring is highly sensitive to instruction design, including the formulation of prompts and the structure of analytic rubrics (Abdullah et al., 2025; García-Varela et al., 2025; Yavuz et al., 2025; Jukiewicz, 2024; Jukiewicz & Wyrwa, 2026). Detailed and structured criteria contribute to more stable and reproducible outcomes, especially when assessment dimensions are clearly operationalised (Yavuz et al., 2025). These findings can be interpreted through the lens of ECD, which conceptualises assessment as a structured relationship between the competencies to be inferred, the tasks used to elicit performance, and the evidence available for scoring (Mislevy et al., 2003). In AI-mediated contexts, this relationship becomes particularly salient, as LLMs operate primarily on the available evidence, processing observable traces and applying scoring decisions based on explicitly defined criteria. As noted by Hao (2025), the integration of AI into assessment highlights how scoring processes depend on the construction and accessibility of evidence, while also raising questions related to core measurement values such as validity, reliability, and fairness.

Consequently, AI-assisted assessment does not function independently of pedagogical design, but rather depends on the coherence between task design, data representation, and assessment criteria (Abdullah et al., 2025; García-Varela et al., 2025; Jukiewicz & Wyrwa, 2026). The model's performance is therefore contingent on the quality and explicitness of the evidentiary framework within which it operates (Hao, 2025; Mislevy et al., 2003). This perspective reinforces the idea that automated assessment systems should be understood as tools whose outputs are shaped by the structure and limitations of the assessment design itself (Voskoboinik et al., 2025; Xia et al., 2024).

2.2 Human and AI-based assessment

A growing number of studies have compared human assessment with AI-generated assessment in educational contexts. The literature converges on the existence of discrepancies between the two, although their direction and magnitude vary depending on factors such as prompt design, task type, and rubric structure (Abdullah et al., 2025; Jukiewicz, 2024; Jukiewicz & Wyrwa, 2026). Importantly, several studies argue that these differences should be interpreted in relation to variability already observed

among human raters. For example, Flodén (2025) found that discrepancies between human and AI scores were comparable to inter-rater variability, suggesting that divergence is not exceptional but inherent to the interpretative nature of assessment. From this perspective, agreement between assessment systems is better understood in terms of overall consistency rather than exact score equivalence (Alderson et al., 1995).

At the same time, a substantial body of research highlights conditions under which AI-generated scores align more closely with human judgement. Specifically, studies consistently show that the use of explicit, well-defined criteria and structured rubrics improves agreement and reduces variability (Abdullah et al., 2025; García-Varela et al., 2025; Yavuz et al., 2025). Conversely, AI performance appears more unstable when assessment criteria are implicit, loosely defined, or highly interpretative. Prompt formulation has also been identified as a critical factor, as even minor variations in instructions can influence how models interpret criteria and generate scores (Ferran, 2025; Jukiewicz, 2024; Jukiewicz & Wyrwa, 2026). These findings suggest that the alignment between human and AI assessment is not intrinsic to the model itself but depends on the conditions under which assessment is structured.

Despite these areas of convergence, important differences remain. Research indicates that AI systems tend to perform more consistently in tasks involving relative judgement, such as ranking, than in tasks requiring absolute score assignment. For instance, Zubaer et al. (2025) found that GPT-4 achieved performance comparable to human raters in ranking tasks, while showing greater variability in point-based scoring. This distinction suggests that AI systems may capture relative differences in performance more reliably than they reproduce absolute scoring scales. In addition, agreement between human and AI assessment is not uniform across assessment dimensions. Studies show higher levels of agreement for criteria that are rule-based and explicitly defined, such as grammatical accuracy, whereas more interpretative dimensions, including content development and discourse organisation, tend to produce greater variability (Abdullah et al., 2025; Yavuz et al., 2025).

In this context, the distinction between reliability and validity becomes particularly relevant. While AI-generated scores may show high levels of consistency and correlation with human ratings, this does not necessarily imply that all dimensions of the construct are equally captured (Jukiewicz, 2024; Yavuz et al., 2025). Rather, alignment appears to depend on the extent to which assessment criteria can be clearly operationalised and applied to observable data (Abdullah et al., 2025; García-Varela et al., 2025). This suggests that similarities between human and AI-based assessment may coexist with systematic differences linked to the nature of the criteria and the type of evidence available for assessment (Voskoboinik et al., 2025; Xia et al., 2024).

Finally, it is important to note that most existing research has focused on written production. The assessment of oral performance remains comparatively underexplored, despite its inherent complexity (de Jong et al., 2025; Luoma, 2011; Ochs, 1979). Oral competence involves multiple dimensions, including prosodic, phonological, and interactional features, which are more difficult to standardise and may not be fully accessible to current AI-based approaches (de Jong et al., 2025; Voskoboinik et al.,

2025). This gap highlights the need for further research examining how AI-based assessment operates in contexts where the available evidence differs substantially from written language.

3. Assessment of oral production using GenAI

The assessment of oral production through the utilisation of GenAI predominantly relies on the transcription of speech into text. In contemporary systems, this process involves two main stages: automatic speech recognition, which converts the acoustic signal into a textual representation, followed by the analysis of this text by LLMs to generate proficiency scores (Louw et al., 2025; Openai, 2024; Voskoboinik et al., 2025). As a result, the assessment is not based on the original oral performance, but on a transformed representation of it. This reliance on transcription introduces methodological constraints that directly affect the nature of the evidence available for assessment. Transcripts do not constitute a complete representation of oral production, but rather a selective and reduced version of the original speech (Ochs, 1979). Prosodic and paralinguistic features such as intonation, rhythm, and pause management are frequently absent or only partially represented. In line with this limitation, Wolf et al. (2023) demonstrate that, although much prosodic information is predictable from text, prosodic features cannot be fully predicted from textual input, which implies that transcripts omit part of the communicative information present in the speech signal.

These constraints have implications for the construct being assessed. Oral competence encompasses multiple interrelated dimensions, including linguistic, phonological, discourse, and interactional features (Coste, 1978; Luoma, 2011). When assessment relies exclusively on transcripts, the available evidence is restricted to textually encoded features, which leads to a partial representation of the construct. Dimensions such as fluency, pronunciation, and interactional management are therefore underrepresented or excluded when only transcript-based evidence is available (de Jong et al., 2025; Voskoboinik et al., 2025).

In response to these limitations, some studies have explored multimodal approaches that incorporate additional indicators of oral performance, such as acoustic features and non-verbal behaviour (e.g. facial expressions, body gestures) (Gaballo, 2024a, 2024b). These approaches aim to extend the range of observable evidence beyond textual data. However, current systems still face difficulties in interpreting higher-level discourse features, such as coherence in interaction, pragmatic appropriateness, or communicative effectiveness.

Despite these constraints, transcript-based analysis has shown potential for assessing specific dimensions of oral production. LLMs have demonstrated the ability to identify grammatical accuracy, lexical use, and aspects of discourse organisation in transcripts, achieving scoring performances that can approximate human judgements under conditions where assessment criteria are explicitly defined (Louw et al., 2025; Voskoboinik et al., 2025).

4. Research aim

The present study examines the validity of ChatGPT as a tool for assessing oral production in higher education. It focuses on three main aspects:

- **Score differences**: analysing the existence, direction, and magnitude of differences between ChatGPT and human ratings
- **Agreement**: examining the degree of alignment between both systems in terms of score distribution and relative ranking of learners
- **Approximation to human judgement**: evaluating whether ChatGPT can approximate human assessment despite potential systematic bias in absolute scoring

The analysis is based on transcript-based assessment using shared analytic criteria, with the aim of determining whether ChatGPT can function as a reliable support tool in the assessment of oral production.

Method

The present study adopted a quantitative comparative design to examine the validity of ChatGPT as a tool for assessing oral production. The analysis focused on differences and alignment between AI-generated scores and human ratings based on shared analytic criteria applied to transcript-based student productions. The design involved a structured workflow including data collection, transcription, dual assessment procedures, and statistical comparison. It is conceptually aligned with ECD approaches to assessment, as it explicitly considers the relationship between the assessed construct, the available evidence, and the criteria used for scoring (Mislevy et al., 2003; Hao, 2025).

4.1 Participants and data

The study analysed 26 oral productions from first-year undergraduate students enrolled in the Bachelor of Teaching and Learning at the University of Andorra. Participants had an approximate A2 level in French according to the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). All productions were anonymised prior to analysis. Each production was assessed by two human raters and by ChatGPT, generating paired data that allowed direct comparison between AI-generated and human scores.

4.2 Assessment instrument

Oral productions were assessed using an analytic rating scale derived from CEFR dimensions. For the purposes of comparison with AI-based assessment, the analysis was restricted to four criteria that could be reliably assessed from transcripts: content and development, coherence and organisation of discourse, syntactic control, and grammatical accuracy. These four criteria were shared by both human raters and the AI system to ensure methodological alignment.

Each criterion was operationalised through three indicators and rated on a 0–3 scale, resulting in a maximum score of 12 points. This restriction reflects the limitations of transcript-based analysis, as only textually observable features were considered (see Additional file 1).

4.3 Procedure workflow

4.3.1 Data collection and oral task

Data collection for the A2 productions analysed in this study took place on 14 and 17 November 2025. The oral task was administered through a custom GPT created by the researcher within the ChatGPT Plus web interface.

Students received a link to access the custom GPT and completed the task individually in a new conversation. Memory was disabled throughout the procedure, and each interaction was conducted separately in order to prevent access to previous student productions. The custom GPT was configured to act as a virtual examiner and to follow a standardised three-part structure: a brief personal presentation, the development of a given topic, and a simulated everyday-life situation (see Additional file 2). All interactions were recorded. At this stage, the custom GPT was used only to administer the oral task and collect the productions; assessment was conducted later through a separate AI-based scoring procedure.

4.3.2 Transcription

Transcription was conducted during the second half of February 2026 through the ChatGPT web interface using a private ChatGPT Plus account. The system operated with the GPT-5.2-series model available in ChatGPT at that time. The same transcription prompt was used for all recordings.

The audio recordings were not directly uploaded to ChatGPT. Instead, anonymised recordings were played for transcription purposes, ensuring that no identifying information about the students was provided. Given the possibility of transcription errors, all transcripts were manually reviewed by the researcher and compared with the original audio recordings to ensure the closest possible correspondence with the spoken productions.

Corrections were limited to transcription accuracy. No grammatical, lexical, or syntactic corrections were introduced, and learner errors, repetitions, hesitations, and reformulations were preserved as far as possible. The final anonymised and manually reviewed transcripts constituted the textual data used for AI-based assessment, while human raters assessed the original audio recordings.

4.3.3 AI-based assessment

Automated assessment was conducted between 8 and 10 March 2026 through the ChatGPT web interface using a private ChatGPT Plus account. The model selected in the interface was GPT-5.3 Instant, the GPT-5-series model available in ChatGPT at that time. Memory was disabled throughout the procedure.

The assessment was conducted through the web interface. Since the ChatGPT interface does not allow full control over model parameters such as temperature or random seed, the full set of anonymised transcripts was assessed three times in three separate conversations using the same standardised prompt to mitigate the influence of single-output variability. In each conversation, the model evaluated all transcripts according to the four selected criteria. The final score for each production corresponded to the mean of the three outputs obtained across the three independent assessment conversations. To estimate intra-model scoring stability, variability across the three ChatGPT assessments was examined by calculating the mean within-case standard deviation and the mean within-case score range across repetitions.

The use of ChatGPT in task administration, transcription, and scoring corresponded to separate stages of the workflow. During task administration, ChatGPT was used as a virtual examiner to elicit oral productions; during transcription, it was used to support the production of verbatim transcripts; and during scoring, it was used as an assessment tool applied only to anonymised and manually reviewed transcripts. The scoring stage was conducted only after anonymisation and manual verification of the transcripts against the recordings. No scoring prompt or analytic rubric was provided to the system during task administration or transcription.

4.3.4 Human assessment

Two experienced language raters independently assessed the oral productions based on the original audio recordings. The assessment was conducted using the same four analytic criteria as those applied in the AI-based assessment, ensuring comparability between both systems. Raters did not have access to each other’s scores.

Agreement between the two human raters was examined using a predefined discrepancy threshold. A total score difference of up to 2 points was considered acceptable, provided that no individual criterion differed by 2 points or more. The discrepancy threshold was used as a descriptive indicator of agreement between the two human raters, not as an adjudication procedure. Cases exceeding this threshold were identified but retained without adjudication, because the study aimed to preserve the original scores assigned by each rater for subsequent comparative analyses with GenAI. The mean of the two human ratings was used as the reference score for the main comparison with GenAI, while the individual rater scores were retained for complementary analyses.

4.3.5 Ethical and data protection considerations

The study was approved by the Academic Board of the UdA. At the beginning of the study, all participating students signed an informed consent form authorising the use of their oral productions for research purposes.

All data were anonymised prior to analysis. Before the transcription stage, the audio recordings were checked by the researcher to ensure that no directly identifying personal information would be made accessible to ChatGPT during transcription. The audio files were not uploaded to ChatGPT as files; however, the audio signal was played to ChatGPT to support the initial transcription process. The resulting transcripts were then manually reviewed by the researcher to ensure fidelity to the original recordings and to remove directly identifying information. For AI-based assessment, only anonymised and manually reviewed transcripts were submitted to ChatGPT. No directly identifying personal data were included in the transcripts used for automated scoring. The use of ChatGPT was therefore restricted to anonymised textual data for the assessment stage, within the framework approved for the study.

4.4. Statistical analysis and validation

AI scores were compared with the mean of the two human ratings using paired observations. The analysis examined:

- differences in scores (direction and magnitude)
- statistical significance (Wilcoxon signed-rank test)
- association between rankings (Spearman correlation)

Normality was tested using Shapiro–Wilk. A paired t-test was computed as a complementary analysis.

4.5 Validation framework and study procedure

Figure 1 presents the relationship between data modality, construct representation, and analytical procedures. The upper section illustrates the validation framework, highlighting the reduction of oral competence to text-accessible dimensions and the resulting methodological constraints. The lower section describes the study procedure, including task design, data collection, transcription, dual assessment, data structuring, and statistical analysis.

Results

The analyses were structured to address three dimensions: score differences, agreement, and ranking consistency. This section presents the comparison between GenAI scores, and the mean scores assigned by two human raters (N = 26). For each case, the difference was calculated as:

Difference = GenAI score – mean human rater score

5.1. Human rater agreement

Before comparing GenAI scores with human ratings, agreement between the two human raters was examined using the predefined discrepancy threshold. For the four criteria retained in the comparative analysis, differences between the two human raters remained within the acceptable range in 22 out of 26 cases (84.6%). Four cases (15.4%) exceeded the threshold. These cases were retained without adjudication in order to preserve the original human scoring variation, and the mean of the two human ratings was used as the reference score for the main comparison with GenAI. The same proportion of discrepant cases was observed in the full audio-based assessment, where four cases also exceeded the predefined threshold of 3 points on the total score out of 24.

5.2 Distribution of score differences

The distribution of score differences was examined using the Shapiro–Wilk test. The result did not indicate a statistically significant deviation from normality ($W = .966$, $p = .519$). Visual inspection showed that most differences were located on the negative side, indicating a tendency for GenAI to assign lower scores than the mean score assigned by the two human raters (Figure 2). However, the distribution did not provide evidence of a significant departure from normality.

Histogram of the score differences calculated as GenAI score minus the mean score assigned by the two human raters ($N = 26$). Negative values indicate lower scores assigned by GenAI. The dashed vertical line represents the mean difference. Although most differences are located on the negative side, the Shapiro–Wilk test did not indicate a statistically significant deviation from normality ($W = .966$, $p = .519$).

5.3 Association between GenAI and human rater scores

The association between GenAI scores and the mean scores assigned by the two human raters was examined using Spearman's rank correlation. The results showed a very strong positive correlation ($\rho = .953$, $p < .001$), indicating a high degree of consistency in the relative ranking of student performance. This suggests that, although differences were observed in absolute scores, GenAI and human raters tended to order students in a highly similar way.

5.4 Difference between GenAI and human scores

A Wilcoxon signed-rank test was conducted as a robust paired comparison between GenAI scores and the mean scores assigned by the two human raters. The results revealed a statistically significant difference between the two assessment systems ($V = 36$, $p < .001$). The mean difference was −1.02 points, indicating that GenAI assigned lower scores on average than human raters. This pattern is

reflected in the individual comparisons, where most observations show lower scores for GenAI (Figure 3).

Each line represents one student and connects the GenAI score with the mean score assigned by the two human raters. Red points indicate GenAI scores, while blue points indicate mean human rater scores. In 21 of the 26 cases, the line rises from GenAI to the human mean, indicating lower scores assigned by GenAI.

5.5 Direction of the score differences

GenAI assigned lower scores in 21 cases (80.8%) and higher scores in 5 cases (19.2%). No identical scores were observed (Table 1). This distribution indicates a predominance of lower scores assigned by GenAI compared with the mean scores assigned by the two human raters.

Table 1. Direction of score differences between GenAI and human raters

Direction of difference	n	Percentage
GenAI higher	5	19.2%
GenAI lower	21	80.8%
Equal	0	0.0%
Total	26	100%

Legend. Differences are calculated as GenAI score minus the mean human rater score.

5.6 Intra-model stability of GenAI scores

Variability across the three ChatGPT assessments was examined to estimate intra-model scoring stability. The mean within-case standard deviation was 0.75 points, and the mean within-case range was 1.40 points. In 9 of the 26 cases (34.6%), the range between the highest and lowest ChatGPT score was below 1 point; in 11 cases (42.3%), it was equal to or below 1 point. These results indicate that some output variation remained across repeated assessments and justify the use of the mean of three outputs to limit dependence on a single model response.

5.7 Relationship between scores

The relationship between GenAI scores and the mean scores assigned by the two human raters is shown in Figure 4. Most observations fall below the line of perfect agreement, indicating that GenAI generally assigned lower scores than the human mean. Despite this systematic difference, the overall pattern shows a very strong positive association, consistent with the Spearman correlation observed between

the two measures (ρ = .953, p < .001). Some dispersion is visible across the score range, indicating that the magnitude of disagreement was not identical across all productions.

Each point represents one student. The dashed diagonal line indicates perfect agreement between GenAI scores and the mean scores assigned by the two human raters. Points below the line indicate cases in which GenAI assigned a lower score than the human mean. The overall pattern shows a very strong positive association between both measures.

5.8 Agreement analysis (Bland–Altman)

A Bland–Altman analysis was conducted to examine the agreement between GenAI and the mean scores assigned by the two human raters (Figure 5). The mean difference was −1.02 points, indicating that GenAI tended to assign lower scores than the human mean. The limits of agreement ranged from −3.40 to 1.35. Most observations fell within these limits, although the magnitude of disagreement varied across individual productions.

Bland–Altman plot comparing GenAI scores with the mean scores assigned by the two human raters. The y-axis shows the difference between GenAI and mean human rater scores, while the x-axis shows the mean of both measures. The solid red line represents the mean bias (−1.02), and the dashed red lines represent the limits of agreement (−3.40 to 1.35). Negative values indicate lower scores assigned by GenAI.

Discussion, implications and future research direction

6.1 Systematic differences between human and AI-generated scores

The results indicate a systematic difference between AI-generated and human scores. The Wilcoxon signed-rank test revealed a statistically significant difference (V = 36, p < .001), with the GenAI system tending to assign lower scores in most cases than the mean of the human raters. This pattern suggests a tendency towards lower scoring in AI-based assessment. This finding is in line with recent research on automated assessment using LLMs, where systematic discrepancies between human and AI scores are observed but often remain within the range of variation documented among human raters (Flodén, 2025). Studies also show that AI-generated scores vary as a function of task type, prompt formulation, and rubric specificity, and that even minor changes in instructions can affect how models interpret assessment criteria (Ferran, 2025; Jukiewicz, 2024; Jukiewicz & Wyrwa, 2026). A growing body of work further indicates that prompt design, rubric structure, and system configuration play a central role in shaping scoring outcomes (García-Varela et al., 2025; Voskoboinik et al., 2025), suggesting that discrepancies between human and AI-based assessments are not random but reflect differences in how assessment criteria are operationalised and processed (Xia et al., 2024). In this respect, the distinction

between reliability and validity is particularly relevant: while AI-based systems may produce consistent outputs, this consistency does not guarantee that all dimensions of the construct are equally represented (Jukiewicz, 2024; Yavuz et al., 2025).

In the present study, one plausible interpretation of the mean difference of approximately 1.02 points is that the GenAI system may have applied the rubric criteria in a more uniform or literal way than human raters. However, the present data do not allow this explanation to be confirmed directly. In contrast to human raters, whose judgement may incorporate contextual interpretation and pedagogical flexibility, the model may have relied more strongly on explicit textual indicators and rubric descriptors (Ferran, 2025; Jukiewicz & Wyrwa, 2026). This may partly explain the tendency towards lower scores in most cases, particularly in productions containing errors or limited development. This was observed in 21 of the 26 productions (80.8%). From an assessment perspective, this difference does not necessarily invalidate the approach, but limits the scope of the validity claim; rather, it reflects differences in how assessment criteria are interpreted and applied. As noted by Xia et al. (2024), the integration of GenAI in assessment involves a shift towards more formalised and explicit representations of assessment processes. Within this framework, the observed difference remains of a similar order of magnitude to the variability reported among human raters in comparable contexts (Flodén, 2025).

These results address the first research objective by confirming the existence, direction, and magnitude of systematic differences between AI and human scoring.

6.2 Association in student ranking

Despite differences in absolute scoring, the results show a very strong association between AI-generated scores and mean human rater scores ($\rho = .953$). This indicates that both assessment systems tend to rank students in a highly similar way. Students who receive higher scores from human raters also tend to obtain higher scores from the GenAI system, and the same pattern is observed for lower-performing students. This convergence in relative ranking constitutes an important indicator of score consistency, though it does not by itself guarantee full construct coverage. In educational assessment, the ability to reproduce relative ordering is considered a key property of automated scoring systems, particularly in comparative contexts (Jukiewicz, 2024; Jukiewicz & Wyrwa, 2026; Zubaer et al., 2025).

However, agreement in ranking should not be interpreted as evidence of full construct equivalence. Previous research has shown that AI systems may achieve high levels of consistency while still varying across specific assessment dimensions, especially those requiring interpretative judgement (García-Varela et al., 2025; Xia et al., 2024; Yavuz et al., 2025). Criteria that are more rule-based and explicitly defined, such as grammatical accuracy, tend to produce higher agreement, whereas more interpretative dimensions, such as content development or discourse organisation, show greater variability (Abdullah et al., 2025; Yavuz et al., 2025).

These findings are consistent with research on automated assessment of open-ended tasks. García-Varela et al. (2025) reported high levels of correlation between human and AI scoring when assessment is guided by structured rubrics. Similarly, Abdullah et al. (2025) observed substantial agreement between human and ChatGPT-based grading. Zubaer et al. (2025) further showed that GPT-4 performs comparably to human examiners in ranking tasks, while differences become more pronounced in point-based scoring. In line with these studies, the present results suggest that ChatGPT can approximate human ranking patterns in terms of relative performance, particularly when assessment is structured and criteria are clearly defined. The model appears to capture text-accessible indicators of performance, such as grammatical accuracy, coherence, and discourse organisation especially for language-related criteria (e.g. grammatical accuracy and aspects of expression), whereas content-related dimensions remain more challenging (García-Varela et al., 2025; Yavuz et al., 2025).

From an applied perspective, this indicates that AI-based assessment may be particularly useful in contexts where the objective is to support comparative interpretation, such as diagnostic assessment or progress monitoring (Jukiewicz, 2024; Zubaer et al., 2025). However, its use remains dependent on the integration of human judgement within a structured assessment framework (Jukiewicz & Wyrwa, 2026; Xia et al., 2024).

This finding directly addresses the second objective by demonstrating a very high level of association in learner ranking despite differences in absolute scores.

6.3 Factors explaining observed discrepancies

The coexistence of systematic differences in scoring and very strong association in ranking suggests that alignment between AI-based systems and human raters should be interpreted as partial and potentially dimension-dependent. Previous research has shown that even when global agreement is high, discrepancies may emerge at the level of specific criteria, particularly those requiring contextual or qualitative interpretation (Abdullah et al., 2025; García-Varela et al., 2025; Yavuz et al., 2025).

In the present study, a first explanatory factor relates to the nature of the data used for assessment. Human raters assessed the original audio recordings, whereas ChatGPT assessed anonymised textual transcripts that had been manually reviewed by the researcher. This methodological distinction is central to the interpretation of the results, since score differences should be interpreted in relation to the specific evidence available to each assessment system. The statistically significant score difference ($V = 36$, $p < .001$), with a mean difference of −1.02 points and lower ChatGPT scores in 21 of the 26 cases (80.8%), may therefore reflect not only differences between human and GenAI scoring, but also differences in the evidentiary basis of assessment. As transcription provides only a selective representation of speech (Ochs, 1979), relevant acoustic and prosodic information, including rhythm, intonation, fluency, pauses, pronunciation, and communicative effort, may be reduced or absent from the textual data available to the model (de Jong et al., 2025; Wolf et al., 2023). This distinction does not invalidate the comparison, but clarifies its scope: the study evaluates the capacity of ChatGPT to assess

text-accessible dimensions of oral production, rather than oral competence in its full acoustic and interactional complexity.

A second factor concerns the way in which LLMs process assessment instructions. Previous research indicates that model performance is highly sensitive to prompt design and task configuration (Ferran, 2025; Jukiewicz, 2024). Even minor variations in instructions can influence how criteria are interpreted and applied, leading to differences in scoring behaviour (García-Varela et al., 2025; Zubaer et al., 2025). In this study, a meta-prompt was designed to standardise assessment and approximate the analytic judgement of a DELF A2 examiner. However, even under controlled conditions, differences remain between human and AI-based interpretation. Human raters may rely on contextual information, communicative intent, and implicit assessment criteria derived from experience, whereas the model is restricted to explicit textual evidence.

This difference in both available evidence and interpretative processes may also be examined through an ECD perspective. From this viewpoint, assessment is understood as a structured relationship between the construct to be measured, the tasks used to elicit performance, and the evidence on which scoring decisions are based (Carton & Ghyselen, 2024; Mislevy et al., 2003). In the present study, the evidence available to the AI system was limited to textual transcripts, whereas human raters had access to the full acoustic signal. This reduction of the broader construct of oral competence to text-based evidence available to the AI system. As noted by Xia et al. (2024), AI-based assessment relies on formalised and explicit representations of assessment criteria, which cannot fully capture the complexity and flexibility of human judgement.

Together, these findings contribute to the third objective by clarifying the conditions under which AI-based assessment can approximate human human scoring for text-accessible dimensions of oral production.

6.4 Implications for the assessment of oral production

Implications for teachers

The results indicate that, although GenAI can reproduce consistent patterns in the assessment of textual features, its outputs remain constrained by the nature of the data on which they rely. The tendency to assign lower scores, combined with the very strong association in ranking, suggests that AI-based assessment can support the consistent application of explicit criteria, but does not replace human interpretation. In this study, the model operated exclusively on transcripts, which limited the assessment to text-accessible dimensions. As a result, aspects of oral competence that depend on prosodic, phonological, and interactional features remained outside the scope of automated assessment (de Jong et al., 2025; Wolf et al., 2023). This reinforces the need for human rater involvement in interpreting performance beyond what can be captured through textual analysis. From this perspective, the role of

the teacher shifts towards the supervision and regulation of AI-supported processes. While GenAI may contribute to consistency and efficiency when structured rubrics are applied (Abdullah et al., 2025; García-Varela et al., 2025), its effective use depends on the teacher's capacity to interpret outputs, identify limitations, and contextualise results in relation to the broader construct of oral competence (Xia et al., 2024).

Implications for assessment design

The findings have direct implications for the design of automated assessment tools. The observed combination of systematic but moderate score differences and very strong association in learner ranking suggests that AI-based assessment may be more interpretable and consistent when criteria are explicit, operationalised, and restricted to observable textual evidence. This result confirms that the performance of LLM-based assessment systems is closely linked to the structure of the assessment framework, including the definition of criteria, the formulation of prompts, and the standardisation of scoring procedures (Abdullah et al., 2025; García-Varela et al., 2025).

At the same time, the study highlights the need to explicitly account for the relationship between data modality and construct coverage. Because the assessment relied on transcripts, only a partial representation of oral competence could be assessed. This supports the view that assessment design must distinguish between dimensions that can be validly assessed through text and those that require access to the acoustic signal (de Jong et al., 2025; Ochs, 1979). In this context, an ECD perspective provides a useful framework for structuring AI-based assessment. It requires a clear articulation between the construct to be assessed, the tasks used to elicit performance, and the observable evidence on which scoring decisions are based (Carton & Ghyselen, 2024; Mislevy et al., 2003). The results of this study illustrate that when the available evidence is restricted, as occurs when oral performance is reduced to textual data, systematic differences in scoring may emerge. Consequently, the design of GenAI-based assessment tools should involve the explicit definition of constructs, the identification of observable indicators, and the alignment between data, criteria, and scoring procedures. In this sense, AI functions not as an autonomous evaluator, but as a system whose outputs depend on the evidentiary structure within which it operates.

Implications for institutional policy

The results also have implications at the institutional level. The combination of very strong ranking association and systematic but moderate score differences suggests that GenAI may support scalable assessment processes, particularly in contexts with large student cohorts (Abdullah et al., 2025). However, the restriction to transcript-based assessment highlights limitations in construct coverage, particularly for oral competence. This indicates that institutional adoption of GenAI should not be limited to efficiency considerations, but must also address issues of validity, transparency, and fairness. In particular, the opacity of automated scoring procedures may hinder the interpretation of results by both

teachers and students (de Jong et al., 2025). From this perspective, institutions are required to define clear frameworks regulating the use of GenAI in assessment, including guidelines on data use, interpretability, and teacher oversight (Chan & Colloton, 2024). The results support the development of hybrid models in which AI-based analysis is combined with human judgement, particularly in domains where multimodal interpretation is required (Wolf et al., 2023; Xia et al., 2024).

6.5 Limitations of the study

The findings should be interpreted in light of several methodological constraints.

First, the relatively small sample size (n = 26) limits the generalisability of the results. Although the paired design allows for direct comparison between AI and human assessment, the findings cannot be readily extended to broader populations without further replication.

Second, the assessment relied exclusively on textual transcripts of oral production. This implies a reduction of the original communicative signal, as prosodic, phonological, and interactional features are not fully represented in text (de Jong et al., 2025; Ochs, 1979). This constraint affects the scope of the construct, as the AI-based assessment only captured the text-accessible dimensions of oral production that were being assessed. This may contribute to the observed differences between AI and human scoring.

Third, the study focused on a single proficiency level (CEFR A2). At this level, oral production remains limited in complexity, which may influence both human and AI assessment. It cannot be assumed that similar patterns would be observed at higher proficiency levels, where linguistic and pragmatic demands are more complex.

Finally, although the use of a standardised meta-prompt and repeated assessments helped estimate and mitigate the influence of single-output variability, LLMs remain sensitive to prompt design and contextual factors (Ferran, 2025; Jukiewicz & Wyrwa, 2026). The observed variation across repeated GenAI assessments also indicates that output stability remains a methodological issue. In addition, the ChatGPT web interface did not allow direct control over parameters such as temperature or random seed. These factors constitute an inherent limitation of LLM-based assessment systems.

6.6 Future research directions

Future research may extend these findings along several lines.

First, replication studies with larger samples and diverse educational contexts are required to assess the robustness of the results. Including learners across different proficiency levels would allow examination of whether the relationship between AI-generated scores and human rating remains stable as linguistic complexity increases.

Second, further work should explore multimodal approaches that combine textual and acoustic data. Integrating prosodic, phonological, and temporal features may provide a more complete representation of oral competence and reduce the limitations associated with transcript-based analysis (de Jong et al., 2025).

Third, additional research is needed on the role of prompt design and rubric structure in shaping AI-generated scores. Systematic comparisons of prompt configurations and scoring protocols may contribute to improving both reliability and interpretability as well as output stability across repeated assessments (Abdullah et al., 2025; Jukiewicz & Wyrwa, 2026).

Finally, future studies should examine hybrid assessment models that combine automated analysis and human judgement. Such approaches may help balance efficiency and interpretative depth, particularly in the assessment of complex and multimodal performances (Xia et al., 2024).

Conclusion

This study examined the validity of ChatGPT for assessing oral production in a higher education context through a comparison with human ratings based on shared analytic criteria. The results show a systematic difference in scoring, with the GenAI system assigning lower scores in most cases, alongside a very strong association in the relative ranking of learners.

These findings contribute empirical evidence to the use of GenAI in oral assessment, indicating that ChatGPT can approximate human scoring patterns for text-accessible dimensions of oral production when assessment is restricted to clearly defined and text-accessible criteria. However, this approximation remains partial. The reliance on transcript-based analysis limits the assessment to textual dimensions and excludes key aspects of oral competence, such as prosody and interaction.

From an applied perspective, the results support the use of ChatGPT as a complementary tool within hybrid assessment models. Its role appears particularly relevant for supporting consistent scoring and learner ranking, while human judgement remains necessary to interpret dimensions of performance that depend on multimodal and contextual information.

Abbreviations

AI: Artificial intelligence

CEFR: Common European Framework of Reference for Languages

DELF: Diplôme d’études en langue française

ECD: Evidence-centered design

GenAI: Generative artificial intelligence

LLMs: Large language models

References

1. Abdullah, S., AlShamy, A., & Al-Harthi, A. (2025, September). A comparative study on the consistency of GPT-based AI grading using human-developed assessment criteria. In *2025 11th International Conference on Education and Technology (ICET)* (pp. 116–120). IEEE. https://doi.org/10.1109/ICET67257.2025.11291030
2. Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.
3. Audran, J. (2024). Cinq enjeux d'évaluation face à l'émergence des IA génératives en éducation. *Mesure et évaluation en éducation, 47*(1), 6–26. https://doi.org/10.7202/1114564ar
4. Carton, C., & Ghyselen, A. (2024). Intégrer l'IA générative dans la conception des évaluations. *HAL*. https://hal.science/hal-05165710
5. Chan, C. K. Y., & Colloton, T. (2024). *Generative AI in higher education: The ChatGPT effect* (1st ed.). Routledge. https://doi.org/10.4324/9781003459026
6. Coste, D. (1978). Lecture et compétence de communication. *Le Français dans le monde, 141*, 25–33.
7. Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.
8. de Jong, N. H., Raaijmakers, S., & Tigelaar, D. (2025). Developing high-quality, practical, and ethical automated L2 speaking assessments. *System, 123*, Article 103796. https://doi.org/10.1016/j.system.2025.103796
9. Ferran, C. (2025). Inteligencia artificial generativa para calificar trabajos académicos. In *ISLA 2025 proceedings* (Paper 19). AIS. https://aisel.aisnet.org/isla2025/19
10. Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal, 51*, 201–224. https://doi.org/10.1002/berj.4069
11. Gaballo, V. (2024, November). Revolutionising language teaching: AI in oral language assessment [Conference presentation]. *Innovation in Language Learning (ICT4LL17)*, Florence, Italy.
12. Gaballo, V. (2024). Innovation in multimodal language teaching and learning: AI-powered oral performance assessment. *RILA: Rassegna Italiana di Linguistica Applicata, 2/3*, 247–267. https://digital.casalini.it/10.1400/301426
13. García-Varela, F., Nussbaum, M., Mendoza, M., Martínez-Troncoso, C., & Bekerman, Z. (2025). ChatGPT as a stable and fair tool for automated essay scoring. *Education Sciences, 15*(8), Article 946. https://doi.org/10.3390/educsci15080946
14. Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings* [Unpublished manuscript].

15. Hao, J. (2025). *AI for educational assessment: Opportunities and challenges through the lens of evidence-centred design and measurement values*. OSF https://doi.org/10.35542/osf.io/dacxj_v2
16. Jukiewicz, M. (2024). The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process. *Thinking Skills and Creativity, 52*, Article 101522. https://doi.org/10.1016/j.tsc.2024.101522
17. Jukiewicz, M., & Wyrwa, M. (2026). Can ChatGPT replace the teacher in assessment? A review of research on the use of large language models in grading and providing feedback. *Applied Sciences, 16*(2), Article 680. https://doi.org/10.3390/app16020680
18. Louw, R., Sharratt, E., de Wet, F., Jacobs, C., Smith, A., & Kamper, H. (2025). *Automatically assessing oral narratives of Afrikaans and isiXhosa children* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2507.13205
19. Luo, J., Keung, C. P. C., & Tang, H. H. (2025). Assessment as a dilemmatic space in the GenAI age: Mapping and unpacking university teachers' conflicting priorities in assessment. *Assessment & Evaluation in Higher Education, 50*(4), 607–621. https://doi.org/10.1080/02602938.2024.2444890
20. Luoma, S. (2011). *Assessing speaking*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733017
21. Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centred design. *ETS Research Report Series, 2003*(1), i–29. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x
22. Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open, 5*, Article 100151. https://doi.org/10.1016/j.caeo.2023.100151
23. Ochs, E. (1979). Transcription as theory. In E. Ochs & B. B. Schieffelin (Eds.), *Developmental pragmatics* (pp. 43–72). Academic Press.
24. OpenAI. (2024, March 20). *Speech-to-text guide*. OpenAI API documentation. https://developers.openai.com/api/docs/guides/speech-to-text
25. Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). Evaluation of the e-rater® scoring engine for the TOEFL® independent and integrated prompts. *ETS Research Report Series, 2012*(1), i–51. https://doi.org/10.1002/j.2333-8504.2012.tb02288.x
26. Voskoboinik, E., von Zansen, A., Phan, N. C., Getman, Y., Grósz, T., & Kurimo, M. (2025). Enhancing second language speaking assessment: Integrating large language models for Finnish and Finland Swedish proficiency scoring. *Language Testing, 42*(4), 508–538. https://doi.org/10.1177/02655322251351648
27. Wolf, L., Pimentel, T., Fedorenko, E., Cotterell, R., Warstadt, A., Wilcox, E., & Regev, T. (2023). Quantifying the redundancy between prosody and text. *arXiv*. https://doi.org/10.48550/arXiv.2311.17233
28. Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education, 21*, Article 40. https://doi.org/10.1186/s41239-024-00468-z

29. Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology, 56*, 150–166. https://doi.org/10.1111/bjet.13494
30. Zubaer, A. A., Granitzer, M., Geschwind, S., Graf Lambsdorff, J., & Voss, D. (2025). GPT-4 shows comparable performance to human examiners in ranking open-text answers. *Scientific Reports, 15*, Article 35045. https://doi.org/10.1038/s41598-025-21572-8

Figures

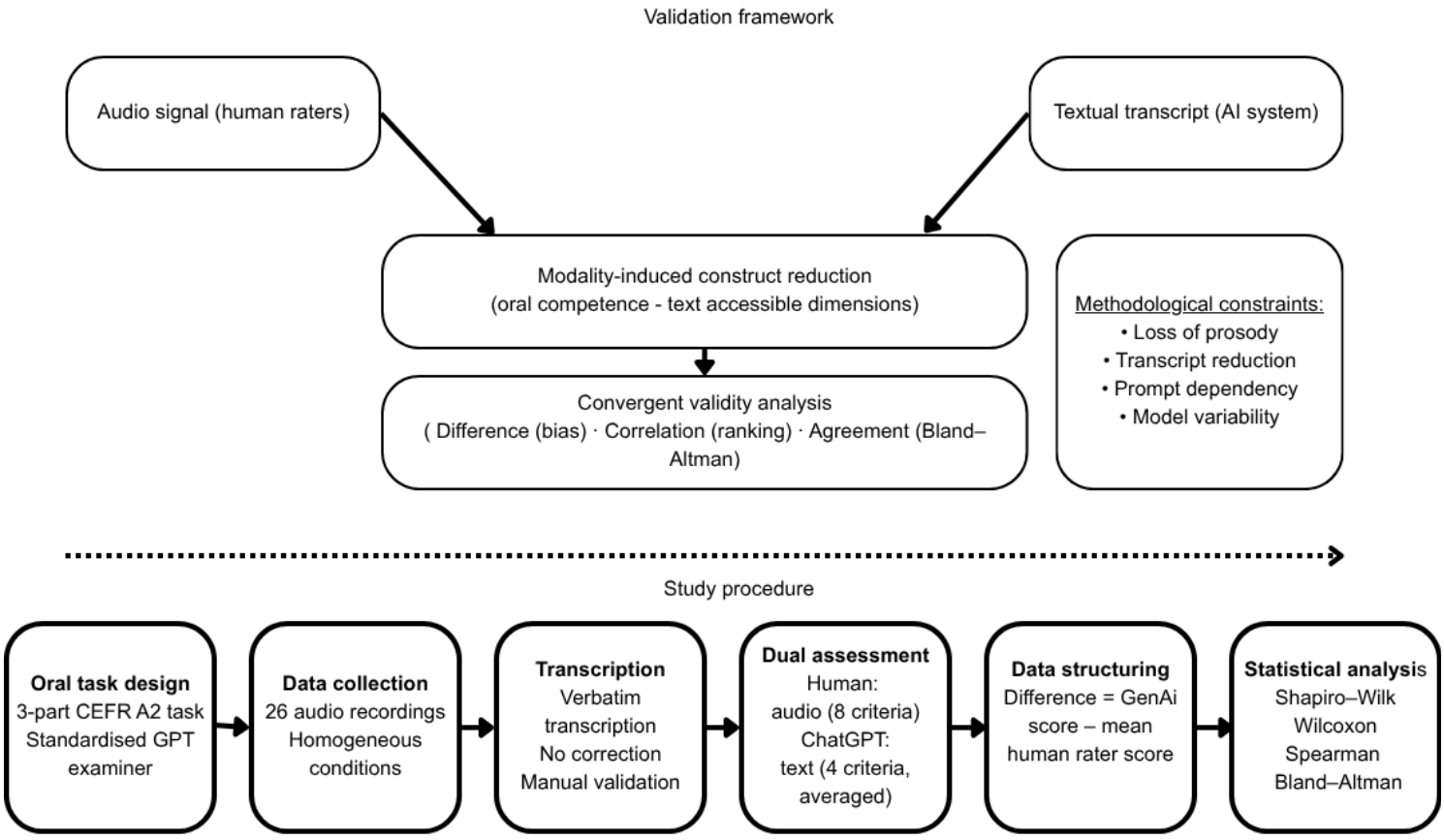


Figure 1

Validation framework and study procedure developed for the present study

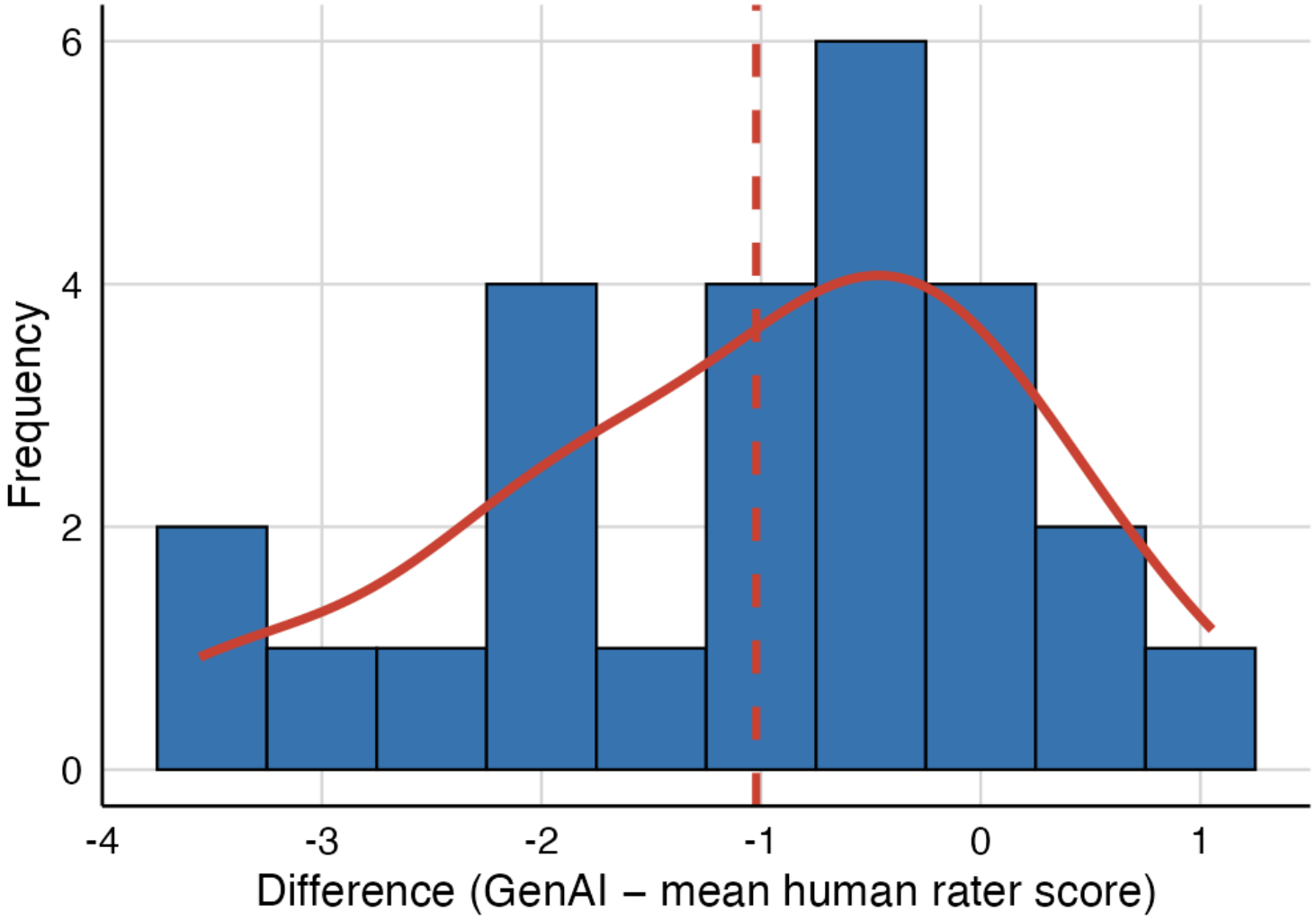


Figure 2

Distribution of score differences

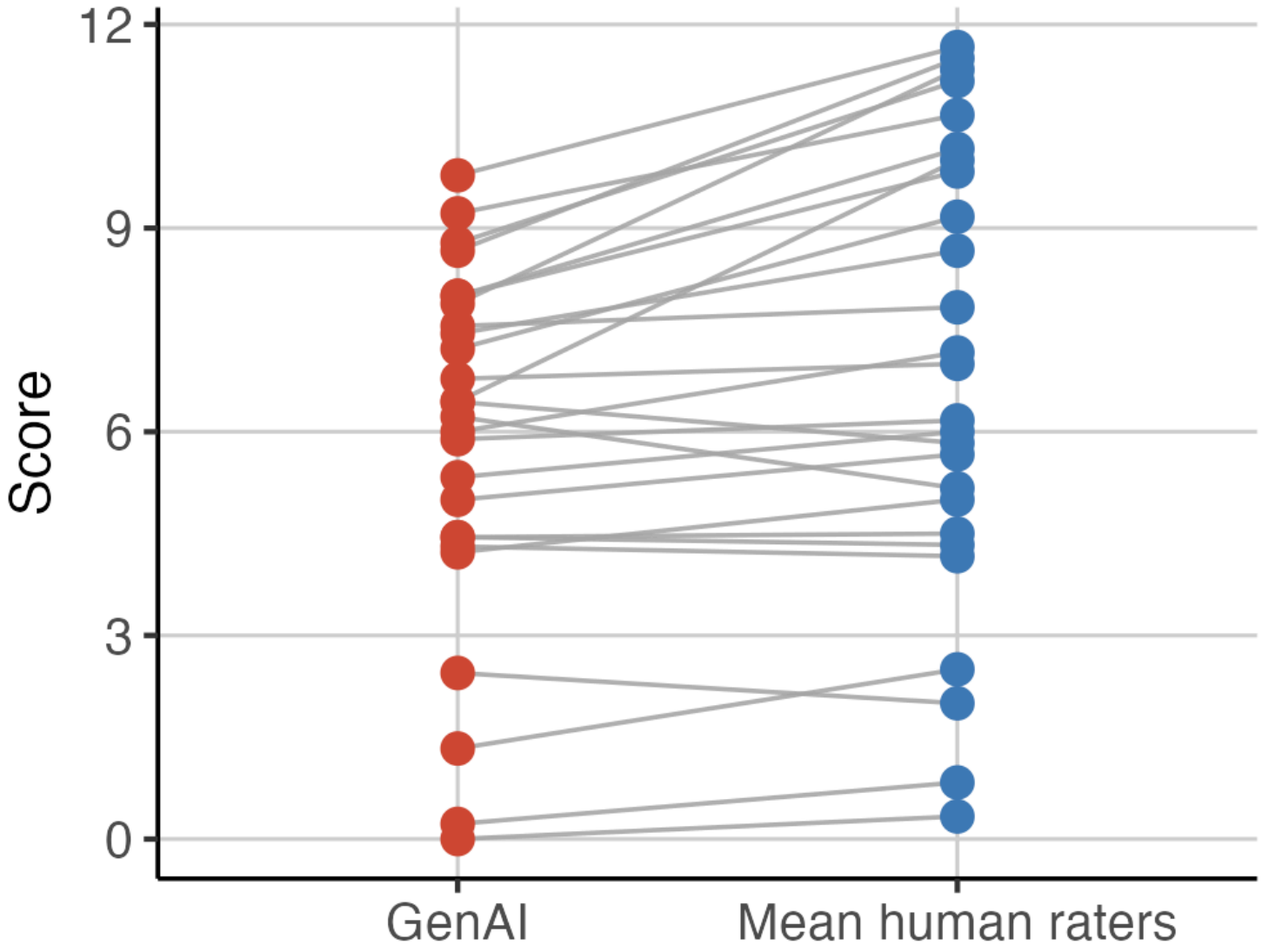


Figure 3

Paired comparison of GenAI and mean human rater scores.

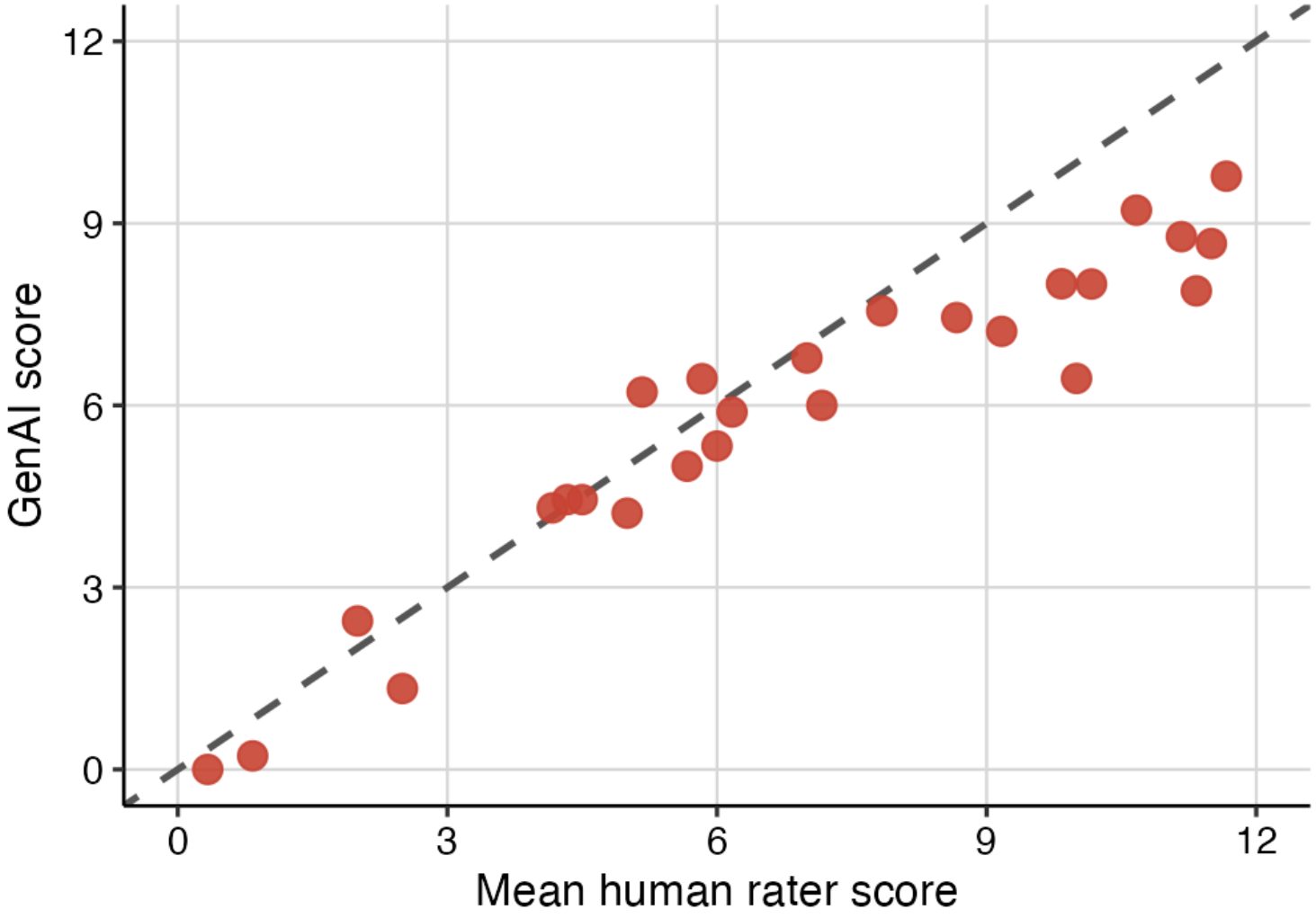


Figure 4

Relationship between GenAI and mean human raters scores.

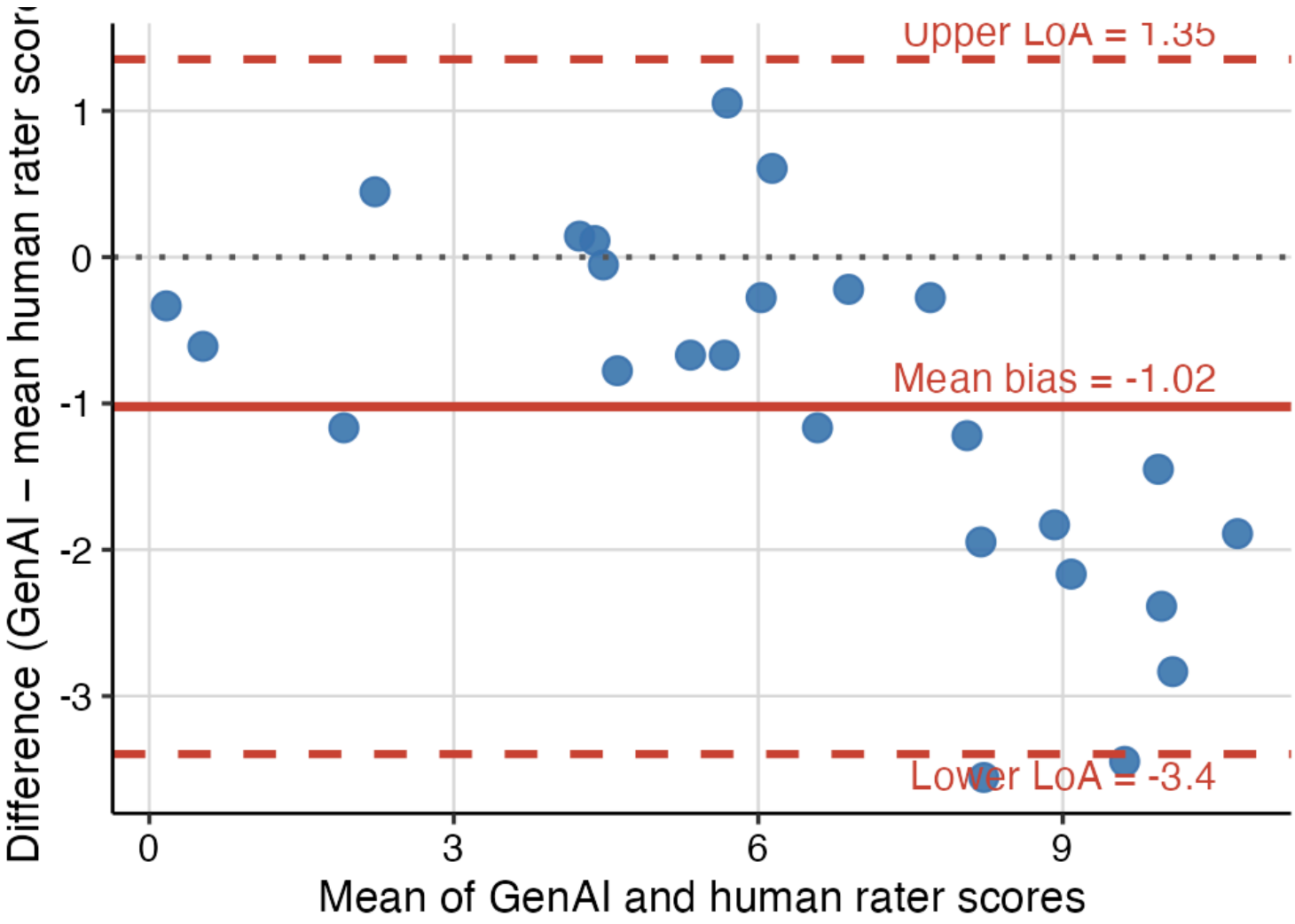


Figure 5

Bland–Altman analysis of GenAI and human raters scores.