International Journal of Language, Translation and Intercultural Communication

Vol 11 (2026)

The Sustainability of Interpreting as a Profession in the Era of Artificial Intelligence


# Assessing Interpreting Performance Through Human and AI Evaluation: Validity, Reliability, and Pedagogical Implications

*Effrossyni Fragkou*

doi: 10.12681/ijltic.45426



**To cite this article:**

Fragkou, E. (2026). Assessing Interpreting Performance Through Human and AI Evaluation: Validity, Reliability, and Pedagogical Implications. *International Journal of Language, Translation and Intercultural Communication*, *11*, 114–153. https://doi.org/10.12681/ijltic.45426

# Assessing Interpreting Performance Through Human and AI Evaluation: Validity, Reliability, and Pedagogical Implications

Effrossyni Fragkou
National & Kapodistrian University of Athens
effiefragkou@enl.uoa.gr

**Abstract**

*This paper examines the use of artificial intelligence (AI) in assessing student interpreting performance in examination settings. It draws on a corpus of 30 audio recordings produced by six students in a first-year healthcare interpreting course within an MA in Conference Interpreting in Canada. The tasks include sight translation (EN<>FR), consecutive interpreting (EN<>FR), and bidirectional medical dialogue in healthcare settings. Student renditions are compared with original source texts, both audio and written, and evaluated against a pre-established assessment grid. The study compares human instructor assessment with AI-based assessment at two points: December 2024-January 2025, during the mid-term examination period, and February 2026, introducing a longitudinal dimension. Using a mixed-methods comparative design, it combines quantitative analysis of scoring patterns with qualitative analysis of convergences and divergences, focusing on accuracy, omissions, additions, distortions, and related assessment criteria. Findings suggest that human assessment better captures prosodic and interactional features, including pronunciation, intonation, rhythm, pausing, speaker attitude, pragmatic force, and hesitation. AI assessment appears relatively stronger in evaluating linguistic and textual dimensions, including content transfer, completeness, grammar, terminology, coherence, and cohesion. The paper also addresses anonymization, voice identifiability, AI use, validity, reliability, and bias.*

*Keywords: Artificial Intelligence (AI), AI-assisted assessment, interpreter assessment, student interpreting performance, healthcare interpreting, sight translation, consecutive interpreting, bidirectional dialogue interpreting, reliability, validity and practicability.*

## 1 Introduction and stakeholder impact

Assessing student performance is central to interpreting pedagogy, with significant repercussions for a wide range of stakeholders. For students, assessment shapes their learning trajectory and determines professional prospects. For the instructor-assessor (rater), it validates the credibility of the instruments and methods used, their readability, availability and usability. At the institutional level, especially within highly competitive international environments, assessment standards reflect the prestige of the training program and its host institution. Furthermore, there is a critical correlation between student performance, degree conferral, and professional accreditation. Finally, for researchers, the pursuit of more efficient evaluation methods is essential to reducing subjectivity and rater bias.

## 2 Challenges in evaluation and study context

Evaluating student-interpreter performance under exam conditions is inherently complex due to the fleeting nature of spoken language translation. This complexity is compounded by technical variables (e.g., recording quality) and pedagogical factors, including the relationship between the student's working languages and the directionality of the task. Additionally, the physical, emotional, and technological circumstances of the exam environment play a decisive role. Notably,

assessment criteria may differ considerably depending on whether the examination is conducted onsite or via online platforms.

This paper draws from an experiment conducted during the mid-term evaluation of six MA students in a North American Conference Interpreting Program. The study explores the potential of technology for interpreting assessment in synchronous online settings (via Zoom) within a hybrid curriculum. The primary objective is to bridge the divide between "low-tech" scoring—emanating from human assessment—and "high-tech" scoring produced by AI-assisted tools. Unlike existing research, which relies on sophisticated computational measures or complex automated performance metrics, this study evaluates the quality of spoken renditions by comparing them to the source text using readily available AI tools.

## 3 Literature review: The evolution of interpreting assessment

Human rater scoring has long occupied a central position in interpreting pedagogy and assessment. Han and Lu (2021) classify the principal approaches to human scoring into five methodological frameworks: (a) atomistic scoring, which focuses on the identification and tallying of errors; (b) questionnaire-based scoring, typically operationalized through Likert-type rating scales; (c) multi-methods scoring; (d) rubric scoring, whether holistic or analytic; and, (e) rank-ordering. Although these approaches draw on the pedagogical expertise and professional judgment of trained assessors and are informed by the nature of the assessment itself (summative assessment for midterm or end-of-year exams vs. formative assessment for feedback during training and practice), they have also been criticized for their vulnerability to subjectivity, inconsistency, and rater bias. In response to these limitations, Han and Lu (2021) advocate a shift toward models of human-machine collaboration, whereby machines ensure speedy, metric-based feedback and routine tasks while humans remain in the loop for contextual diagnosis, ethical oversight, and interpretative judgment. Such collaboration may enhance both the psychometric reliability and the social accountability of interpreting assessment (fairness, transparency, defensibility of assessment and answerability to all stakeholders).

### 3.1 The shift toward Automatic Machine Scoring (AMS)

To mitigate limitations associated with exclusively human evaluation, a research has turned to automatic machine scoring based on predetermined computational procedures for calculating quality indicators. According to Han and Lu (2021), this technologically driven approach has evolved along four principal trajectories.

The first trajectory concerns *temporal variables*, whereby acoustic features such as pause duration, articulation rate, and speech rate are extracted through specialized software to estimate perceived fluency. The second focuses on *linguistic surface features*, using Natural Language Processing (NLP) techniques to quantify variables such as word count, syntactic complexity, and grammatical accuracy. A third line of development involves the application of *machine translation (MT) metrics*, including BLEU, NIST, METEOR, and TER, which compare student output with one or more human reference translations. Finally, *quality estimation* (QE) approaches employ machine learning algorithms trained on human-annotated data in order to predict interpreting quality even in the absence of a reference text. Collectively, these approaches reflect an important shift toward scalable and potentially more standardized forms of assessment.

More broadly, automated assessment research in interpreting has so far been dominated by computationally intensive approaches that depend on quality-estimation pipelines, supervised machine-learning models, engineered fluency features, or neural MT-based metrics. This trend indicates that the move away from purely human scoring has not simply involved the adoption of digital tools, but rather the growing reliance on technically sophisticated models designed to approximate or predict expert evaluative judgment.

### 3.2 Neural Metrics and Predictive Modeling

More recent scholarship has increasingly emphasized the potential of neural-based metrics to outperform traditional n-gram-based measures in the assessment of interpreting performance. Han, Lu, and Chen (2025) report that metrics such as BLEURT-20 and COMET-22 display strong positive correlations with human ratings, particularly with respect to fidelity and accuracy. Alongside these developments, researchers have also designed integrated predictive models intended to approximate human evaluative judgment.

Several studies illustrate this methodological shift. Stewart et al. (2018) adapted MT quality estimation to interpreter output, thereby extending quality-estimation logic to the assessment of simultaneous interpreting performance. Wang and Yuan (2023) combined delivery- and fidelity-related features in supervised scoring models, while Wang and Wang (2024) identified acoustic-temporal predictors for machine-based fluency assessment. Related work has also drawn heavily on MT evaluation metrics: Macháček et al. (2023) examined the performance of BLEU, chrF2, BERTScore, and COMET against human ratings in simultaneous speech translation, and Han and Lu (2025) demonstrated the potential of neural MT metrics for the large-scale automatic assessment of English-Chinese interpreting. Taken together, these studies show that automated interpreting assessment is increasingly shaped by predictive modeling and metric-driven evaluation procedures that require substantial technical design, feature selection, and, in many cases, specialized computational expertise.

Similarly, Jiang and Zhang (2025) demonstrate that XGBoost can achieve particularly strong predictive performance for fluency through the analysis of "breakdown" features, including the frequency of filled pauses. In a similar vein, Wang et al. (2023) employ Support Vector Machines (SVM) to combine delivery-related and information-based variables, attaining a prediction accuracy of 62.96% for "pass" outcomes. These findings suggest that AI-based systems may have considerable utility, particularly in the preliminary screening of large-scale interpreting assessments.

### 3.3 The frontier of LLMs and Explainable AI (XAI)

The emergence of Large Language Models (LLMs), such as GPT and Claude, marks a further stage in the technological development of interpreting assessment. Wang and Wang (2025) found that Claude demonstrates a relatively strong correlation with human evaluation. At the same time, however, they identify what they term a "style gap", whereby LLMs tend to privilege formal written-like expression and linguistic sophistication, rather than the communicative effectiveness, reformulation strategies, and oral mediation skills that are central to interpreting performance.

In parallel with the growing use of LLMs, efforts have also been made to address the opacity often associated with automated scoring systems. Jiang and Zhang (2025), for instance, incorporate Explainable AI (XAI) through SHAP (SHaply Additive Explanation) analysis to generate more

transparent and interpretable feedback in an attempt to open the black box. Their findings indicate that neural embedding-based metrics such as BLEURT are among the most powerful predictors of information completeness, while Chinese-specific phraseological diversity indicators, such as CN_RATIO, play an important role in predicting target-language quality. Such work is especially significant because it attempts to reconcile predictive performance with interpretability, thereby enhancing the pedagogical usefulness of AI-assisted assessment.

### 3.4 Research gap and the current study

Notwithstanding these technological advances, an important research gap persists with regard to the identification of the most effective combination of analytical tools, ranging from conventional linguistic indicators to newer LLM-based evaluative systems, for generating robust and pedagogically meaningful assessment outcomes across diverse interpreting classroom settings. Existing literature points to at least three major challenges in this regard.

First, the "black box" problem remains unresolved, insofar as increasing model complexity often diminishes the transparency and diagnostic value of automated assessment for educational purposes. As Alon and Levkovich (2026: 7) observe there are consistent reports that GenAI is a black box for users because outputs are visible, but it is difficult to inspect how these outputs are produced and what would be the consequences of their failure. Second, the literature presents conflicting evidence concerning the relative importance of particular features across different dimensions of interpreting quality. Third, as shown above, many existing studies rely on highly sophisticated automated metrics and large-scale annotated datasets that are typically unavailable to individual instructors or institutions and difficult to operationalize or even conceptualize before implementing them in routine teaching practice.

It is precisely at this point that the present study departs from the dominant trend in the literature. Whereas much of the existing research relies on computationally intensive approaches such as quality-estimation pipelines, supervised machine-learning models, engineered fluency features, or neural MT-based metrics, the current study explores a more accessible alternative. In contrast to the advanced computational procedures employed in studies such as Jiang and Zhang (2025), Han and Lu (2025), Stewart et al. (2018), Wang and Yuan (2023), Wang and Wang (2024), and Macháček et al. (2023), the present research uses readily available AI tools to compare spoken renditions directly with the source text (audio and/or script). In doing so, it prioritizes methodological accessibility and practical replicability over metric sophistication.

More specifically, the study examines the interaction between low-tech scoring, understood as human assessment, and high-tech scoring through the use of mainstream AI tools and/or LLMs (such as ChatGPT) that do not require specialized programming expertise, extensive corpora, or custom-built computational pipelines. The aim is to investigate whether off-the-self platforms can offer a viable assessment bridge for instructors operating in highly competitive and hybrid educational environments, where time constrains, accuracy, reliability and fairness requirements must be perfectly balanced. By directly comparing student renditions with source texts (audios and/or scripts) through broadly available technological means, this study seeks to contribute to the democratization of automated interpreting assessment within a human-in-the-loop framework and to reposition such practices from the domain of advanced data science to that of the everyday interpreting classroom.

## 4 Methodology

### 4.1 Research Design

The study uses a longitudinal embedded mixed-methods comparative design to examine the extent to which AI-assisted assessment aligns with human assessment of MA student interpreting performance in examination settings. The design is longitudinal in that it compares AI-based assessment conducted at two different points in time. The first assessment occurs during the original examination period in December 2024-January 2025 (Time 1). The subsequent assessment takes place in February 2026 (Time 2), using the same student performances and the same assessment grid. The study is comparative because it contrasts human assessment with AI-based assessment, while comparing AI assessment at Time 1 with AI assessment at Time 2. A subsequent human assessment occurs at Time 2 to contrast it with the one produced at Time 1 and against AI assessments at both Time 1 and Time 2. Finally, this is a mixed-methods design because it combines quantitative analysis of scoring patterns with qualitative analysis of recurrent convergences and divergences between human and AI ratings.

The analysis unfolds in three interrelated laywers. The first layer relates to the quality of different transcripts produced from the student recordings to determine whether speech-to-text output is sufficiently accurate for assessment purposes. The second layer consists in comparing student rendition transcripts to the original source texts (audio and scripts) so as to identify transfer-related features such as omissions, additions, distortions, and completeness of information. The third layer aims at comparing AI-generated assessment and human assessment on the basis of a pre-established evaluation grid. In this sense, transcript fidelity is treated not merely as a technical preprocessing step, but as a methodological variable that may affect the quality and validity of AI-assisted assessment.

The study is case-based oriented. Each recorded student performance constitutes a bounded case, while each case contains several embedded units of analysis, including *task type* (consecutive interpreting, dialogue interpreting and sight translation), *language direction* (English into French, French into English and bidirectional), *transcript version* (depending on the tools used), *assessment mode* (tools used), and *assessment moment* (Time 1 and Time 2). This design makes it possible to investigate not only whether human and AI assessment differ, but also where, how, and under what conditions those differences emerge and how they can affect overall assessment.

### 4.2 Research context, participants, and corpus

The dataset derives from the 2024-2025 cohort of MA students enrolled in a Conference Interpreting programme at a Canadian university. All participants belong to the English<>French language stream and are in the first year of study, during which they complete a mandatory healthcare interpreting course. The English<>French cohort comprises six students with the following language profiles: four with Canadian French as their A language, one with American English as their A language, and one fully bilingual student (Canadian English and Canadian French)[1].

[1] Working language classification is based on the entry exam students had to sit to be admitted in the program and reflects the classification used by AIIC, EMCI and EU interpreting services.

Demographic information relating to participant self-identification was collected solely for descriptive purposes and was not used as a basis for participant labelling in the analysis. The cohort included four students identifying as women, one identifying as a man, and one identifying as non-binary. All six students, as well as the instructor (female), identified as White/Caucasian. While the present study does not examine whether ASR performance correlates with speaker race, gender, or language, it is possible that some of the findings may reflect characteristics of the English- and French-language acoustic and language models underlying the ASR systems used. In future work, race and gender could be considered as possible stratification variables, in combination with task type, language direction, accent/dialect, and speaker role. Accordingly, future research could consider race and gender as potential stratification variables, alongside task type, language direction, accent or dialect, and speaker role. Age might also be examined in subsequent work, although it was excluded from the present study for specific reasons: in adult education contexts as the one examined here, identifying participants by age may itself be regarded as discriminatory; moreover, the available literature tends either to focus on adult-child differences in language-specific contexts (Feng et al., 2024; Safavi et al., 2018) or to find no clear and systematic bias across adult age groups (Liu et al., 2022). At the same time, current research does not support a uniform relationship between speaker gender and speech-to-text performance. While differential performance across gendered speaker groups has been reported, findings remain inconsistent across studies, suggesting that ASR bias is multifactorial and better understood as an interactional, data-dependent phenomenon rather than a universal one (cf. Feng et al., 2024).

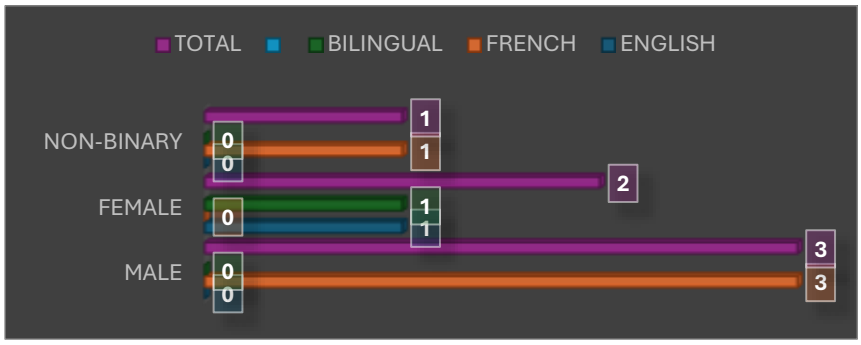


*Image 1: Student distribution per gender and working language*

The corpus of this study consists of 30 audio recordings in total: 24 recordings of student performances in examination-related tasks of a medical nature, and 6 recordings containing the instructor's comments on each student's performance. The student-performance recordings cover three interpreting task types in the English<>French language pair. These include sight translation in both directions (one recording per student, corresponding to two separate sight translation tasks, one into French and one into English), consecutive interpreting in both directions (two recordings per student, one into French and one into English), and a bidirectional medical dialogue interpreting (one recording per student).

The corpus captures authentic student performances produced under actual examination conditions rather than data generated specifically for research purposes. All tasks focuse on the gastrointestinal system. The sight translation task requires students to orally translate sections of an informational fact sheet intended for allophone patients undergoing colonoscopy, with the aim of explaining the procedure and the necessary preparation. The consecutive interpreting task into

French is based on a 3-minute-56-second audio recording of a male intern (gastrointestinal specialist) recounting his first experience of administering a suppository in a hospital setting. The consecutive interpreting task into English consists in a 6-minute-16-second speech on gastrointestinal health delivered by a healthcare professional in a peer-education context. Finally, the bidirectional dialogue task is a 2-minute-56-second simulated medical consultation involving a specialist, a female adolescent patient, and the patient's parents, during which a diagnosis of Crohn's disease is communicated. These tasks are selected because they reflect key modes of interpreter training in healthcare-related contexts and differ in that they foreground meaning transfer, interaction management, delivery, and prosodic performance. The selection of topic, task length, level of difficulty, and assessed skills are designed to correspond to the students' acquired knowledge and expected level of competence at this stage of the curriculum (mid-term), including short sight translation tasks, consecutive interpreting tasks of up to six minutes, and brief bidirectional dialogues based on topics previously addressed in class. Taken together, the tasks mirror the kinds of communicative demands and professional expectations healthcare interpreters are required to meet in real-world practice.

This corpus strengthens the ecological validity of the study, as all performances are elicited under genuine pedagogical and evaluative conditions. At the same time, the limited number of participants and the fact that data derive from a single course classify the study as a small-scale exploratory investigation. Its purpose is, therefore, not to support broad generalization, but rather to generate methodologically robust and contextually grounded findings.

### 4.3 Assessment conditions and instruments

Student recordings as described above are assessed under four conditions:

a) Human assessment at Time 1: assessment conducted by the instructor following the original examination period (December 2024), using the established evaluation grid. The assessment was launched on January 8, 2025.
b) AI assessment at Time 1: assessment conducted using AI tools available during the original assessment period, based on the same performances and the same evaluation grid (January 8 to January 15, 2025).
c) AI assessment at Time 2: assessment conducted in February 2026 using the same recordings, the same source materials, and the same evaluation grid, in order to examine whether changes in AI capability affect alignment with human assessment and whether newer AI tools could yield better results.
d) Human assessment at Time 2: assessment conducted in February 2026 using the results of AI assessment at Time 2 and examine whether such results could influence Human assessment at Time 1, especially in respect to accuracy and completeness.

To the extent possible, the study seeks to maintain procedural consistency across the two AI assessment moments (Time 1 and Time 2) by using the same performance data, the same evaluation criteria, and a standardized prompting procedure, where prompting is required. This standardization is essential in order to attribute any observed differences, at least in part, to the evolution of AI capabilities rather than to changes in task material or assessment instructions.

Assessment is based on a pre-established analytic evaluation grid designed for interpreter performance (see Annex). The grid is organized into three major domains: *Content, Form,* and

*Delivery*. These domains collectively capture the multidimensional construct of interpreting performance, including transfer accuracy, linguistic realization, and oral delivery.

The *Content* domain includes criteria such as accuracy of meaning transfer, completeness of information, correct rendition of main ideas, appropriate transfer of qualifiers and hedges, and avoidance of omissions, additions, distortions, and opposite meaning. The *Form* domain addresses grammaticality, syntax, lexical and terminological appropriateness, coherence, cohesion, and overall quality of linguistic formulation. Finally, the *Delivery* domain includes fluency, pausing, hesitation phenomena, pronunciation, intonation, rhythm, and other performance-related aspects of oral production.

### 4.4 Evaluating the assessment grid as a research instrument

The assessment instrument utilized in this study is best characterized as a weighted analytic rating scale operationalized through discrete checklist-like criteria. It is analytic in that it assesses distinct dimensions of interpreting performance separately—most notably content transfer, target-language form, and delivery—rather than subsuming performance under a single holistic judgment. Its performance-level descriptors further support quality judgments across a continuum, an important advantage of analytic scales (Brookhart, 2018). At the same time, its checklist-like character derives from the fact that each domain is defined through specific observable criteria. This design is consistent with broader scholarship on performance assessment, which emphasizes the value of analytic rubrics for assessing complex constructs. It also aligns with interpreting-assessment research, where interpreting quality is increasingly understood as multidimensional rather than reducible to isolated error counts. (Jonsson & Svingby, 2007; Pöchhacker, 2001; Shafiei, 2024). More generally, analytic descriptive rubrics appear to be preferred in higher education because their descriptive nature is more supportive of learning (Brookhart, 2018: 22–23).

A central strength of the grid is its close alignment with established conceptualizations of interpreting competence. Interpreting quality has long been understood to involve not only fidelity of sense transfer but also cohesion, completeness, fluency, grammar, terminology, delivery and pragmatic efficacy. User-oriented and practitioner-oriented research has shown that interpreting is evaluated through multiple criteria rather than semantic accuracy alone. A grid organized around *Content*, *Form*, and *Delivery* captures the multidimensional nature of interpreting quality more adequately than models based solely on omissions or lexical errors (Kurz, 2001; Pöchhacker, 2001). It also offers clear diagnostic value. By scoring these domains separately, it enables assessors to determine whether weaknesses stem primarily from content transfer, target-language formulation, or oral performance rather than reducing performance into a single comprehensive score.

This is one of the main pedagogical advantages of analytic rubrics in both general assessment research and interpreter education. In the broader rubric literature, Jönsson and Svingby (2007: 140-141) argue that rubrics can produce positive educational effects by making expectations and criteria explicit and thereby facilitating feedback and self-assessment. Panadero and Jönsson (2013: 140), for their part, conclude that rubrics may support improved performance through transparency, feedback, and self-regulation. As far as interpreting assessment is concerned, Shafiei (2024: 2–3) shows that analytic rating evaluates multiple criteria separately, provides detailed feedback, and is particularly useful for identifying aspects of performance that can be addressed through targeted instruction. An instrument structured around *Content*, *Form*, and *Delivery* is especially appropriate when assessment is intended not only to assign a grade but also to render learners' strengths and

weaknesses across component skills visible, discussable and comparable at the individual learner level but also across learners (e.g., within a specific cohort or across cohorts).

The grid also has important practical advantages. Compared with full transcript-based error analysis, it is considerably more feasible in classroom and examination settings because it does not require the assessor to establish an elaborate taxonomy of errors before scoring. It was developed in response to pedagogical demands: the need for a tool that can capture key dimensions of student interpreting performance, support feedback, and remain workable under authentic examination conditions. In this sense, it should be understood as a pragmatic, pedagogically grounded instrument. Its value lies not only in its immediate usefulness for rating but also in its capacity to generate domain-specific data that can inform teaching, learning, and future assessment design by reducing rater bias.

For the all the aforementioned reasons, the grid is well suited to assist in achieving the goals set out in this study: it is multidimensional, transparent, pedagogically meaningful, and operationally feasible, while also enabling comparison between AI and human scores both globally and by domain. This, in turn, allows for a more nuanced analysis of convergence and divergence across assessment modes. At the same time, the instrument is not treated as a finalized or fully validated measurement tool. Like many pedagogically developed assessment instruments, it includes elements that may require refinement, such as greater descriptor specificity, closer calibration of criteria, and possible reconsideration of grouping or weighting particular features. These issues are not external limitations but part of the rationale of the study itself, which subjects the grid to empirical examination across rating contexts and assessment modes. The instrument is thus approached both as a practical scoring tool and as an object of inquiry whose structure may be confirmed, refined, or revised on the basis of empirical evidence. The study therefore contributes not only to the comparison of AI and human scoring, but also to the ongoing validation of a theoretically informed and pedagogically grounded assessment instrument (Han et al., 2024; Shafiei, 2024; Wind, 2020).

### 4.5 Data preparation and transcription procedures

Data preparation was undertaken in three interrelated stages: (a) preparation of the source materials for each examination task, (b) transcription and verification of student performances, and (c) alignment of source texts, recordings, transcripts, and assessment outputs for comparative analysis. This staged procedure was necessary in order to ensure consistency, traceability, and analytical reliability of the corpus used in the study.

#### 4.5.1 Preparation of source materials

The first stage concerned the identification and preparation of the source materials associated with each examination task, namely the source texts and/or source audio recordings. These materials were collected and, where necessary, transcribed, normalized, or converted into written form so as to ensure consistency across tasks and facilitate subsequent comparison with student renditions. Because the three examination components differed in format and mode of delivery, the preparation procedure was adapted to the specific requirements of each task.

For the consecutive interpreting task from English into French, the instructor selected and adapted a source text that would meet the pedagogical and evaluative requirements of the examination. The final script was then imported into Narakeet, an AI-based text-to-speech

voiceover generator, in order to produce the source audio used during the exam. A male English Canadian voice was selected for this purpose. In practice, Narakeet provided only one English Canadian male voice option ("Ryan Male"), which was retained as the most contextually appropriate choice, given that the students were predominantly originated from Canada, were enrolled in a Canadian university and would identify as English Canadian speakers.

For the consecutive interpreting task from French into English, the source material was compiled and adapted from multiple texts so as to incorporate the linguistic, thematic, and terminological features considered necessary for the examination. A script was then produced and recorded in French by the instructor using her own voice. The recording was created in Audacity. As the instructor is not a native speaker of French, she attempted to approximate a metropolitan French accent in the recording, given her educational background.

For the bidirectional dialogue task, the instructor developed a simulated medical consultation scenario drawing on a range of medical and supporting resources. The script was subsequently entered into Narakeet to generate the corresponding audio material. Because this platform cannot process bilingual scripts simultaneously, the English and French utterances were entered separately in two distinct script sets and were then concatenated in the predetermined sequence so as to recreate the dialogue. In the resulting recording, the francophone physician was represented by a Standard French Canadian male voice ("Xavier"), while the patient, a 15-year-old adolescent, and the patient's parents were each assigned separate voices: a young female Standard British English voice for the patient ("Emma"), a mature male Standard British English voice for the father ("Nelson"), and a female Standard British English voice for the mother ("Rosalind").

Once the source materials had been finalized, each student performance was linked to the relevant metadata, including task type, language direction, student identifier, and recording date. This step ensured that all renditions could be systematically traced across tasks, transcription versions, and assessment phases.

### 4.5.2 Transcription of student performances

The second stage involved the transcription of all student recordings. Because the study explicitly examines the use of AI in the assessment of interpreting performance, particular attention was paid to the quality, fidelity, and analytical consequences of speech-to-text output. Multiple transcript versions were compared against the original audio recordings in order to determine the extent to which transcription inaccuracies might affect AI-generated assessment. This is especially important in the case of healthcare interpreting, where minor lexical errors, mistranscribed medical terminology, incorrect speaker attribution, or problems of punctuation or segmentation may affect the representation and interpretation of meaning.

Transcription followed a clearly defined, multi-tool procedure. At Time 1, the recordings were transcribed using the Speech-to-Text tool of the European Commission's AI-based multilingual services, an official component of the Commission's broader AI language technology (eLangTech) environment. This tool belongs primarily to the category of automatic speech recognition (ASR), while also functioning as a captioning and subtitling service and, more broadly, as a task-specific productivity tool for public-sector and institutional use. It allows eligible users, including public administrations, SMEs, academic institutions, NGOs, Digital Europe projects, and EU job candidates, to upload audio or video files and generate transcriptions or subtitles in all official EU languages as well as a number of additional languages.

From the standpoint of data governance, the tool forms part of the Commission's "free and secure AI language tools" environment, meaning that data are processed under strict EU data-protection rules (GDPR) and are not used to train commercial AI models. Its principal advantages lie in its accessibility, institutional orientation, and usefulness for administrative, academic, and professional communication, particularly in contexts such as meetings, workshops, and conferences. At the same time, its limitations are methodologically significant. First, the Commission does not guarantee output accuracy; as a result, human verification remains necessary, particularly in high-stakes contexts. Second, performance may vary substantially depending on elements such as recording quality, accent, speaker overlap, terminology density, and other acoustic conditions. Third, publicly available documentation provides a useful functional overview of the tool but does not offer full technical specifications of parameters such as diarization, timestamping, speaker separation, confidence scoring, export options, or file-processing constraints.

For the purposes of the present study, further limitations proved particularly consequential: all automatically generated transcripts did not systematically capture disfluencies through dedicated transcription conventions or special notation. As a result, important features of student delivery were frequently lost, reduced, or normalized (i.e., 'sanitized') in transcription. These included filled pauses (e.g., *uh, um, er, ehm, emm*), silent pauses of varying duration, mid-clause and between-clause pauses, intra-word pauses, and repair phenomena such as false starts, self-corrections, reformulations, repetitions, substitutions, abandoned utterances, and truncations. Other delivery-related disruptions, including prolongations, broken words, non-lexical vocalizations, irregular tempo, articulatory restarts (false starts), and pronunciation-related breakdowns, were also not consistently preserved if preserved at all. This limitation is methodologically important, since such features are relevant to the assessment of fluency, delivery, processing difficulty, and interactional management in interpreting performance.

At Time 2, additional transcription tools were introduced in order to compare output quality and examine the implications of different transcription environments for downstream assessment. ChatGPT Plus (GPT-5.4 Extended Thinking mode), Gemini 3.1 Pro (Thinking mode), and Notta.ai may be situated within the broader landscape of AI-enabled transcription technologies as proprietary, cloud-based systems rather than transparent standalone ASR engines. In the case of ChatGPT and Gemini, transcription is embedded in multimodal large language models (LLMs), that is, transformer-based architectures trained on large-scale datasets and designed to process and generate language across modalities. These systems incorporate speech-to-text functionality within broader environments that support summarization, translation, and higher-order analysis, whereas Notta.ai is more appropriately classified as a dedicated transcription platform oriented toward recording, speaker labeling, and transcript production. This distinction is methodologically relevant because it frames these tools not as neutral ASR systems but as integrated sociotechnical infrastructures with distinct analytical affordances (OpenAI, n.d.; Google, n.d.; Notta, n.d.).

The tools were selected for two main reasons: they are the most popular and most readily available tools; they purportly represent complementary transcription paradigms. ChatGPT Plus and Gemini Pro, both used here in their paid "thinking" configurations, are characterized by reasoning-oriented architectures that allow intermediate analytical processing prior to output generation. As advertised by the tools themselves, this feature is particularly relevant for tasks requiring contextual interpretation, transcript refinement, and performance analysis. Similarly, because Gemini Pro can offer an extended context window (up to approximately one million

tokens) and integrate with productivity environments such as Google Docs, Gmail, and Slides, it can reportedly process longer recordings and more complex datasets. Notta.ai, by contrast, was selected for its specialized transcription workflow, including speaker identification, segmentation, and export functionalities. All these features facilitate the production of structured transcripts for subsequent human review. The tool was also recommended by ChatGPT and Gemini Google when they failed to yield results as expected.

This selection is also methodologically significant in light of the literature on AI-assisted transcription, which emphasizes that transcription is not a purely mechanical step but an interpretive process shaped by technological mediation (McMullin, 2023). The use of proprietary cloud-based systems introduces additional considerations related to data governance, informed consent, and platform-based processing of potentially sensitive audio data (Da Silva, 2021; Samuel & Wassenaar, 2025). Moreover, research in speech technology demonstrates that voice data may remain identifiable even when transcripts are anonymized, and that ASR performance varies across speakers and linguistic profiles, raising issues of bias and reliability (Nautsch et al., 2019; Koenecke et al., 2020; Feng et al., 2023). Accordingly, the use of these tools in the present study is treated as a methodological choice informed by the context of the experiment. For this reason, it required special attention to platform settings, data-protection safeguards, and systematic human verification of machine-generated transcripts.

In terms of their performance, and against original expectations based on advertised system capabilities, ChatGPT Plus (Extended Thinking Mode 5.4) was unable to produce transcripts given the parameters set out for the purposes of this research (see prompting in Annex). The explanation provided by the tool was that the environment did not provide sufficiently robust native automatic speech recognition (ASR) systems for high-fidelity offline transcriptions. Although all uploaded MP3 files were uncorrupted, accessible, decodable, and a speech-recognition package was available locally, the model files required for transcription were not cached in the environment, and external downloading was not possible. Additional offline recognition pathways were explored, but the available fallback system produced output with substantial lexical distortion and inadequate overall accuracy. Given these constraints, the tool failed to produce verbatim transcripts. Its decision to return no results was methodologically significant. The study required strict verbatim transcription, including dysfluencies, hesitations, false starts, incomplete utterances, and other features of spontaneous speech. Prior versions of ChatGPT (namely 5.3 Instant version) were equally unsuccessful because the requirements set out in the prompt outlined highly granular specifications, which required a functioning audio processing/ASR pipeline with acoustic playback capability unavailable in the execution environment of all aforementioned versions.

Similarly, Gemine Google admitted not possessing the ability to natively listen to the acoustic audio track because its platform uses a background speech-to-text pipeline to process the audio into text before handling it to the tool. This background pipeline is designed for standard readability, which means that all pauses, stutters, false starts and fillers are automatically filtered.

The recordings were also processed using Otter.ai, a specialized transcription platform. This was done in an effort to reinforce the strict-verbatim requirement, since this tool operates more directly at the acoustic level and therefore tends to preserve, rather than normalize, certain disfluencies. However, the results were very poor which led to selecting Notta.ai (paid version). The latter choice proved to be far more reliable as transcripts in both English and French were much closer to the recordings. The tool managed to produce a much higher accuracy level in both languages, despite considerable challenges in terms of the pronunciation of some of the students.

It should be noted that, unlike ChatGPT, and Gemini, the European Commission Speech-to-Text tool and Notta.ai do not offer a prompting function. Consequently, their output is produced independently of any user-defined prompting parameters.

### 4.5.3 Alignment and preparation for comparative assessment

The third stage involved preparing the corpus for comparative human and AI-based assessment. The corpus was first evaluated through human assessment at Time 1, followed by AI-based assessment at Time 1. At Time 2, an additional round of human assessment was conducted to confirm or challenge the initial evaluations in light of the results obtained at Times 1 and 2. The analytical sequence was as follows.

Before processing, all recordings were anonymized by eliminating any trace of the students' first and last name as well as any other personal data that could be inadvertantly revealed. Students were labelled as subjects 1, 2, 3, 4, 5, and 6 respectively, with each number referring to the alphabetical order of their first name. Recordings and their corresponding word texts were then labelled starting with the Time (1 or 2) the assessment took place, the task to be assessed, the language combination followed by the 'transcription' identifier and the number assigned to each student in the context of anonymization. Labelling was as shown below:

- Time 1
  - 2025-01-08 Consec EN-FR_transcription_1
  - 2025-01-08 Consec FR-EN_transcription_1
  - 2025-01-08 Dialogue_transcription_1
  - 2025-01-08 ST 1_transcription_1
- Time 2
  - 2026_02 Consec EN-FR_trancription 1
  - 2026_02 Consec FR-EN_transcription 1
  - 2026_02 Dialogue_transcription_1
  - 2026_02 ST_transcription_1

At the subsequent phase, student rendition transcripts were compared, contrasted, and crosschecked with the corresponding authentic source texts and audio recordings in order to identify transfer-related features central to interpreting assessment. This alignment procedure made it possible, to the extent this was achievable, to distinguish between deviations attributable to each student's interpreted output and distortions introduced by transcription technology limitations. Discrepancies such as omissions, additions, distortions, loss of qualifiers, reduction of hedging, semantic shifts, and other departures from the source message and/or from the recorded student performance were systematically tagged.

The transcripts were then uploaded to ChatGPT Plus and Gemini Pro at Times 1 and 2 of the experiment, together with the corresponding recordings and the evaluation grid. Using the same prompt structure throughout, each generative AI system was asked to generate an assessment report based exclusively on the criteria contained in the evaluation grid for each student and for each

examination component. These AI-generated reports were subsequently compared (Time 2) with human assessment carried out at Times 1 and 2 using the same grid.

### 4.5.4 Ethics of AI-assisted transcription

In this study, transcript fidelity was treated as a methodological issue, not merely a technical one. AI-based assessment did not evaluate student performance directly. Rather, it evaluated a textual representation shaped by audio quality, transcription accuracy, and the preservation or normalization of interactionally important features. In this respect, the reliability of AI-assisted assessment must be interpreted in relation both to the assessment model and to the representational integrity of the transcript on which it relied. This is consistent with literature that treats transcription as a methodologically consequential stage of analysis rather than a neutral clerical procedure (McMullin, 2023; Eftekhari, 2024).

Equally, the use of AI-assisted transcription was approached as an issue of data governance and research ethics, not simply of technical convenience. Recent work has emphasized that automated transcription must be considered on the basis of informed consent, accountability, and the sociotechnical infrastructures (technical infrastructure, data infrastructure, human and organizational layers, governance and regulatory frameworks) through which speech data are processed (Da Silva, 2021; Herdiyanti, 2024; Samuel & Wassenaar, 2025).

This concern was particularly relevant in this study because transcription was conducted through personal paid subscriptions rather than institutionally-governed environments. In the former case, protections are typically structured through provider defaults and user-managed privacy settings. In the latter case, protections are governed by formal academic data-processing agreements. Given that voice recordings are highly information-rich forms of personal data that may function as biometric identifiers and support inferences about identity and other speaker characteristics, anonymization of the resulting transcript cannot fully eliminate privacy risk once raw audio has been uploaded to third-party systems (Nautsch et al., 2019a, 2019b; Kröger et al., 2020; Bäckström, 2025). Moreover, research on privacy-preserving speech technologies indicates that voice de-identification remains technically difficult and involves a persistent trade-off between privacy and utility rather than guaranteeing anonymity (Tomashenko et al., 2022; Srivastava et al., 2022; Leschanowsky et al., 2025).

Given the risks described above, all available platform-level mitigation measures were implemented across ChatGPT Plus, Gemini Pro, and Notta.ai in order to reduce exposure, circulation, and retention of identifiable audios. These measures included disabling model-training and data-sharing options where available, minimizing identifiable information in uploaded files, restricting permissions and access, and, most importantly, deleting recordings and transcripts promptly after processing. Even so, the use of proprietary AI systems was not treated as ethically neutral, since opaque data infrastructures and under-specified data provenance, which go beyond the control of the user, continue to pose challenges for accountability, bias assessment, and meaningful consent (Bender & Friedman, 2018; Bender et al., 2021). These risks are reinforced by emerging concerns about the downstream misuse of voice data. This includes voice cloning and related forms of secondary exploitation (Barnett, 2023; Bélisle-Pipon et al., 2024). They are also reinforced by documented ASR disparities associated with accent, race, gender, age, and non-native speech. Such disparities may affect transcript fidelity and, in turn, the validity of downstream interpretation (Koenecke et al., 2020; Feng et al., 2024).

Voice de-identification (i.e., altering recordings to obscure speaker identity), a procedure that could potentially mitigate some of the aforementioned risks, was not implemented in this study. Such procedures require advanced techniques that were not available at the time of the study. Addtionally, they introduce a substantial risk of signal distortion, with likely negative effects on ASR accuracy and on the preservation of prosodic and interactional features, essential to assessing interpreter performance. This could potentially compromise the validity of the analysis. Prior to participation, students were informed that their recordings would be anonymized and used for research purposes, were given sufficient explanations of the risks involved and the mitigation measures in place, and explicit informed consent was obtained from all participants.

## 5 Data delineation and analysis

### 5.1 General considerations

The study pivoted around a series of interrelated research questions derived from its central objective—to determine the extent to which AI-generated ratings align with human assessment of student interpreting performance. The broader aim was to explore whether readily available AI tools can support interpreter assessment in a manner that is reliably accurate, timely, and methodologically defensible for pedagogical use, without requiring advanced technical expertise on the part of the instructor-rater. More specifically, the analysis examined whether the degree of alignment between human and AI assessment varies according to task type and on the basis of the three principal domains (criteria) of our evaluation grid, that is, *Content*, *Form*, and *Delivery*. It also investigated whether AI-assisted assessment changed over time by comparing outputs produced at Time 1 and Time 2, and by tracing the forms of convergence and divergence that emerged across human assessment, AI assessment, and AI-informed re-examination of student performances. Finally, the study explored the extent to which transcript fidelity conditions the usefulness of AI-assisted assessment and, more specifically, the extent to which machine-generated transcripts can support (or distort for that matter) the rater's own judgment of students' performance.

Addressing these questions required the use of a mixed-methods analytical framework, since the study combined quantitative comparison of scoring patterns with qualitative examination of recurrent divergences and convergences between assessment modes. Such a design is appropriate when the aim is not only to measure differences across conditions, but also to explain how and why such differences arise within a complex evaluative process (Creswell & Plano Clark, 2018). The quantitative component focused on scores resulting from comparing across assessor type, task type, rubric domain, and assessment time (Time 1 and Time 2). For each recording, the data extracted and tabulated included total scores, domain-level scores, criterion-level scores, and, where relevant, observations concerning transcript quality. This made it possible to compare not only global assessment outcomes but also the distribution of agreement and discrepancy across the distinct dimensions of the evaluation grid.

The qualitative component consisted of a criterion-based discrepancy analysis, through which cases of meaningful convergence and divergence between human and AI ratings were identified and examined. The purpose of this analysis was not simply to document differences, but to explain their source in relation to the nature of the task, the criterion evaluated each time, and the affordances and constraints of transcript-mediated AI assessment. In procedural terms, this qualitative layer drew on a structured coding logic akin to thematic analysis, insofar as

discrepancies were systematically categorized and interpreted across the dataset (Braun & Clarke, 2006).

The use of a domain-based evaluation grid was particularly important. As stated earlier in this paper, analytic rubrics are considered especially well suited to complex performance assessment because they render multidimensional constructs assessable by separating them into distinct, observable criteria instead of subsuming performance under a single holistic judgment (Jonsson & Svingby, 2007; Brookhart, 2018). In the present study, this approach was deemed essential because the three main components of the grid—*Content*, *Form*, and *Delivery*—are not equally accessible to transcript-mediated AI evaluation. Certain aspects of content transfer and linguistic form can be more readily inferred from transcripts, whereas delivery-related phenomena such as hesitations, pausing, back-tracking, false starts, intonation, or interactional timing are much more dependent on the audio signal. In this respect, the rubric was not only an assessment instrument but also an analytic device for testing whether AI performs differently across dimensions of interpreting competence. The criterion-based discrepancy analysis further served a formative and diagnostic purpose by revealing where alignment held and where it broke down, which is consistent with broader literature on the value of rubrics in supporting transparent and diagnostically useful assessment (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013; Brookhart, 2018).

#### 5.1.1 Establishing the 'Gold Standard'

A necessary precondition for this analytical sequence was to establish a transcription baseline (also know as 'the golden dataset'). Before student performances could be assessed through AI-assisted procedures, it was first necessary to determine whether the speech-to-text systems used to generate transcripts produced output of sufficient fidelity to support downstream evaluation. For this reason, a benchmark corpus was created. In this validation stage, only two transcription systems were retained for systematic comparison: the Speech-to-Text tool of the European Commission's Digital Europe environment (hereafter Europa) (produced at Time 1 and Time 2 of the experiment with no difference in the transcripts yielded) and Notta.ai (hereafter Notta) produced during Time 2. Both systems were tested initially not on student renditions but on the original recordings and scripts corresponding to the various examination components, namely the consecutive task from English into French, the consecutive task from French into English, and the bidirectional medical dialogue, (benchmark corpus). Sight translation could not be validated as the original was in written form and not an audio. The purpose of this benchmarking phase was to establish a reference point against which the transcription of student performances could later be interpreted.

The logic of this procedure was both methodological and practical. The source-task recordings were expected to yield relatively high ASR accuracy because they had been produced under more controlled acoustic conditions than the student performances themselves. The bidirectional dialogue, for example, relied on AI-generated voices with relatively stable and standard pronunciations, while the consecutive source recordings were delivered at a comfortable rate, with clear diction, and generally satisfactory sound quality. Even if the narrator was not a native speaker, this was not treated as a methodological weakness. On the contrary, since real healthcare encounters routinely involve non-native speakers, such conditions were considered ecologically defensible, even if they potentially increased the cognitive load placed on the students.

To assess ASR fidelity, each Europa and Notta transcript was compared against its corresponding benchmark transcript after the removal of timestamps, speaker labels, boilerplate, and other non-linguistic metadata. Word Error Rate (WER) was calculated as token-level

Levenshtein distance divided by the number of reference words[2], while Character Error Rate (CER) was calculated as normalized character-level Levenshtein distance after orthographic normalization. WER is a widely used metric in automatic speech recognition evaluation, and its use here was intended to provide a transparent, standardized basis for comparing ASR output across files and systems (Morris et al., 2004). However, for the purposes of the study, WER and CER were not considered sufficient on their own. As research on ASR evaluation has shown, overall error rates do not fully reflect the practical consequences of transcription errors, especially when those errors affect meaning or occur at important points of the interactional exchange (Morris et al., 2004; McCowan et al., 2005). For this reason, the quantitative benchmarking in Table 1 was complemented by aligned error windows in Table 2, allowing errors to be examined not only by frequency but also by type, location, and semantic weight.

As Table 1 shows, ASR fidelity proved to be task-dependent across all three benchmark files. In the bidirectional dialogue (on the topic of Crohn disease, hereinafter Crohn speech), Europa performed with near-verbatim accuracy, producing only two substitution errors and yielding a WER of 0.46%, whereas Notta's performance was markedly less stable once the raw-ASR estimate was considered, rising to 8.92% and driven primarily by deletions. This suggests that Notta was substantially less robust especially when dealing with rapid turn alternation, short reactive utterances, and interactionally sensitive dialogue structure. The consecutive task speech from French into English (on the topic of gastrointestinal health, hereinafter GI speech) yielded a different profile and was the most difficult file for both systems. Although Europa produced slightly more total word errors than Notta (34 vs. 30), its output was more substitution-heavy, whereas Notta's was more insertion-heavy. This distinction is important because substitution-heavy output is more likely to compromise semantic and terminological accuracy, whereas insertion-heavy output more often generates transcript noise without necessarily altering propositional content to the same extent. The consecutive English into French speech (on the topic of constipation, hereinafter Constipation speech), by contrast, produced the strongest aggregate results for both systems. Once again, Europa performed very well, with a WER of 0.92%, while Notta matched the control exactly. Taken together, these three files show that ASR performance in the present material cannot be characterized globally; rather, it varies not only by system but also by task format, interactional density, reference quality (i.e. the acoustic quality of the benchmark corpus as such), and pronunciation (see Voice Generation Column).

| File | Voice Generation | System | Ref. words | Hyp. words | S | D | I | Total edits | WER | CER |
|---|---|---|---|---|---|---|---|---|---|---|
| Crohn dialogue | AI-generated | Europa | 437 | 437 | 2 | 0 | 0 | 2 | 0.46% | 0.11% |
| | | Notta (surface-corrected) | 437 | 435 | 7 | 3 | 1 | 11 | 2.52% | 1.02% |
| | | Notta (raw-ASR estimate) | 437 | 406 | 6 | 32 | 1 | 39 | 8.92% | 6.41% |
| GI speech | Human-generated (non-native) | Europa | 751 | 757 | 26 | 1 | 7 | 34 | 4.53% | 1.87% |
| | | Notta | 751 | 764 | 17 | 0 | 13 | 30 | 3.99% | 2.27% |
| Constipation speech | AI-generated | Europa | 543 | 544 | 4 | 0 | 1 | 5 | 0.92% | 0.24% |
| | | Notta | 543 | 543 | 0 | 0 | 0 | 0 | 0.00%* | 0.00%* |

*Table 1: ASR performance across benchmark files: normalized edit operations, WER, and CER*

[2] In our context this means that the ASR transcript was compared to the transcript of the corresponding benchmark corups word by word; then substitutions, deletions and insertions were considered and literaly counted then divided by the number of words in the reference (i.e., original) script.

*Note 1: S = substitutions; D = deletions; I = insertions; WER = word error rate; CER = character error rate.*

*Note 2: The zero-error Notta result for the constipation speech reflects textual dependence between the hypothesis and control file rather than an independent perfect ASR outcome because the original script was uploaded on Notta ai to help with the transcription of the corresponding audio recording.*

Taken together, the results reported in Table 1 and illustrated in Table 2 (below) directly informed the later stages of the study. First, they demonstrated that transcript fidelity cannot be assumed, even under relatively controlled recording conditions, and must therefore be treated as a methodological variable in its own right. Second, they showed that ASR systems fail in different ways, with distinct implications for assessment. Third, and most importantly, they showed that the usefulness of an ASR transcript for assessment depends not simply on the volume of errors, but on their semantic distribution and evaluative weight. For this reason, transcription benchmarking was not treated here as a preliminary technical exercise but as an indispensable building block of the study's methodological architecture. More fundamentally, this supports a broader methodological claim central to this article: the validity of AI-assisted assessment of interpreting performance depends on how that performance is transcribed and represented for analysis.

| File | System | Reference span | Hypothesis span | Error profile | Error Assessment |
|---|---|---|---|---|---|
| Crohn dialogue | Europa | "d'être venus aujourd'hui" | "d'être venu aujourd'hui" | Minor agreement substitution | Non-critical |
| Crohn dialogue | Europa | "vous avez ressenties" | "vous avez ressenti" | Minor agreement substitution | Non-critical |
| Crohn dialogue | Notta | "Mon but est d'aider Emma" | "Mon but est DDMA" | Severe lexical distortion | Critical |
| Crohn dialogue | Notta | "No way! We want a second opinion" | Material partly omitted / compressed | Deletion of turn-initial content | Critical |
| Crohn dialogue | Notta | "I just want her to be okay" | "I just want out to be okay" | Semantic substitution | Almost critical |
| GI speech | Europa | "gastroentérologie" | "gastro-hétérologie" | Terminological distortion | Critical |
| GI speech | Europa | "habitudes intestinales" | "abysses intestinales" | Semantically disruptive substitution | Critical |
| GI speech | Europa | "syndrome du côlon irritable" | "syndrome du collant irritable" | Medical lexical distortion | Almost critical |
| GI speech | Europa | "reflux gastrique" | "refus gastrique" | Clinically risky substitution | Critical |
| GI speech | Notta | "près de 95 %" | "près de quatre-vingt-quinze pour cent" | Numeric expansion / insertion | Non-critical |
| GI speech | Notta | "les problèmes de peau" | "les pro-- les problèmes de peau" | False start / insertion noise | Non-critical |
| Constipation speech | Europa | "What now, doc?" | "What now, dog?" | Lexical substitution (Critical) | Critical |
| Constipation speech | Europa | "past the anosphincter" | "past the anal sphincter" | Terminology normalization / substitution (Non-critical) | Non-critical |

*Table 2: Representative aligned error windows by benchmark file and ASR system*

*Note. Reference span = wording in the control transcript; hypothesis span = corresponding wording in the ASR transcript. Examples are illustrative rather than exhaustive and were selected to show both low-impact and meaning-bearing deviations.*

Table 3 provides a more analytical explanation of the nature of errors per transcript. Focus should be placed on the "Interpretation" column, where one can observe error behavior (omissions, hallucination-like outputs, lexical distorsions, terminological drifts, segmentation noise) against their methodological implications, namely the extent to which such errors may affect speaker turns, diagnostic terms, or key meaning-bearing units. When examined closer, this table goes beyond numbers providing a more analytical interpretation of the types of errors as opposed to the sheer

volume of errors. For example, in the GI speech, lexical substitutions may be more damaging than insertion noise because distorted medical terminology may lead to content misinterpretation. Inversely, omissions as the ones observed in the Grohn dialogue can be particularly problematic because they may negatively affect the interactional structure beyond the word level, and by extension, the communicative purpose of the dialogue. It can be safely deduced that not all transcription errors carry the same degree of consequence. As a result, each table should be read with some caution because the functional seriousness of the errors depicted in the "Interpretation/implication" column suggest that labels such as "less/more stable" or "noisier transcript" are merely mechanical labels (i.e., characterizations under "Dominant deviation profile"); they are also analytical (interpretative) and as such they depend on the judgement as well as on the purpose and usefulness of the transcription, which, in turn, is task-type based. In other words, one should refrain from labelling one system as being better or more reliable than the other. Instead, it would be more appropriate to address errors from the standpoint of vulnerability, with each system having its distinct vulnerability profile. As the pie charts in Images 1 & 2 demonstrate, ASR errors are task-sensitive, acoustic-sensitive, unevenly distributed, and evaluative weighted.

| Transcript (File) | System | Dominant deviation profile | Interpretation / implication |
|---|---|---|---|
| Crohn dialogue | Europa | Very minor morphosyntactic drift | Near-control output; deviations primarily agreement/inflection. |
| Crohn dialogue | Notta | Omissions, hallucination-like output, severe lexical distortion (in annotated file) | Less stable in dialogue/turn-sensitive content; annotated transcript likely understates raw ASR errors. |
| GI speech | Europa | Terminology distortion, lexical substitution, agreement drift | Some errors are semantically high-impact in medical French (terminology substitutions). |
| GI speech | Notta | More insertions/fragments/repairs, segmentation noise | Often preserves meaning but produces noisier transcript (disfluency fragments, insertions). |
| Constipation narration | Europa | Small colloquial/terminology normalization shifts | High overall fidelity; small substitutions (e.g., address terms, anatomy term normalization). |
| Constipation narratio | Notta | Apparent perfect match to control | Not evaluable independently because the control appears Notta-derived (lineage confound). |

*Table 3: Error profile by transcript*

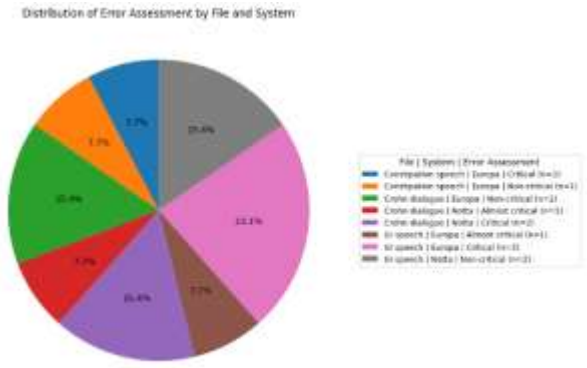


*Image 1: Distribution of error by file and system*

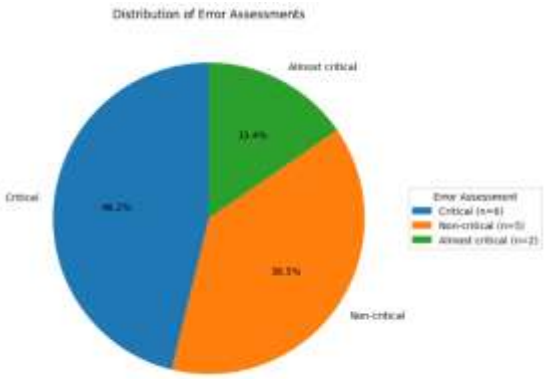


*Image 2: Distributio of error assessments*

### 5.1.2 Analyzing our data

From all six case analyses undertaken for the purposes of this experiment, it became clear that AI-assisted interpreting assessment cannot be evaluated as a simple contest between tools, on the one hand, and humans, on the other. This is particularly important with regard to the first component of the comparison, namely the apparent contest between AI tools. For this reason, the experiment was deliberately designed as a layered pipeline rather than as a direct comparison of isolated systems.

At the first upstream layer, Europa and Notta, that is, speech-to-text systems, were used to produce the transcripts subsequently used in the downstream analysis. These tools did not assess student performance. Rather, they determined which aspects of the students' oral performance became available as textual evidence. At the second downstream layer, ChatGPT and Gemini were used to assess the evidence produced at the first layer by applying the 21-criterion analytic rubric divided into *Content*, *Form*, and *Delivery*. Finally, the combined evidence produced through this pipeline was compared and contrasted with human assessment.

The distinction between these two layers is critical. Transcription tools do not assess; they shape the evidentiary basis on which assessment becomes possible. Assessment tools, for their part, cannot neutrally access student performance unless the audio evidence and transcripts on which they rely are sufficiently reliable. They infer quality from a mediated representation of performance, and the quality of that representation inevitably affects their evaluative output. For this reason, as already argued in this paper, transcript fidelity cannot be treated as a merely clerical or technical preliminary step. It constitutes a methodological variable in its own right, capable of altering downstream scoring.

This distinction is also central to any meaningful model of human-AI coordination. Human assessors are able to integrate audio evidence, task context, pedagogical history, and the communicative weight of errors in ways that automated assessment tools cannot consistently replicate. AI assessment, by contrast, may identify patterns, generate structured feedback, and support cross-case comparison, but it may also amplify transcription errors, introduce hallucinated observations, overstate precision, or produce bias through the very evidentiary chain on which it depends.

Consequently, the analysis presented below does not ask which tool is best in isolation. Instead, it examines where human and AI judgment converge when these tools are used in combination, where they diverge, and how such convergence and divergence vary by student profile, task, domain, criterion, and transcription pathway. The ultimate question is not whether AI can replace human assessment, but whether this layered procedure is pedagogically sound and useful enough to justify its considerable cost in time, cognitive effort for the rater, and quality-control burden, while avoiding the risk that bias, hallucination, and false precision may cloud rather than clarify the rater's judgment.

### 5.1.3 AI-human alignment

Our analysis showed considerable internal variation in the human benchmark scores. The mean human task score is 79.94/100, but the scores range from 66.62 to 93.25 when students are grouped according to profile level, including factors such as A language, B language, or perfect bilingual status. The range is even wider when scores are examined by task and directionality, moving from

57 to 95 depending on whether the task is consecutive interpreting, dialogue interpreting, or sight translation, and whether the language direction is EN-FR or FR-EN. This is an important finding because it shows that AI-human alignment is not being tested against a narrow or uniform group of performances. The pooled data include both stronger and more uneven interpreting profiles. This makes the alignment problem, and the original hypothesis of the study, more realistic.

The best AI-human alignment is useful, but it is not stable enough to justify replacement scoring. The closest AI configuration produces a pooled MAE of 7.67 percentage points, with values ranging from 6.00 to 9.33. The second-closest configuration produces a pooled MAE of 9.61 percentage points, with values ranging from 7.50 to 11.72. These figures suggest that AI output may be helpful for moderation, second reading, or comparison, but they also show that even the strongest configuration may still differ from human task scores by a margin which is pedagogically significant.

More specifically, as will be shown later in this analysis, uniform prompting did not produce uniform assessment behavior at the downstream level. The two assessment tools differed not only in terms of leniency or severity in domain-by-domain scoring, but also in the way they justified their scores. Compared to Gemini, ChatGPT generally provided a far more detailed account for each domain and criterion across tasks. It included verbatim examples to support scoring decisions; summarized the main strengths and weaknesses of each performance in bullet form; identified urgent priorities for improvement; and, ended with a final qualitative verdict ranging from "limited" or "unsatisfactory" to "satisfactory" or "fully effective". These features made its feedback more transparent, relevant and potentially usable from a pedagogical viewpoint. However, transparency should not be confused with reliability. Even detailed AI feedback may still be affected by transcript quality, selective attention, over-interpretation, or inaccurate weighting of errors.

The overall figures also explain why it would be methodologically unsafe to make broad claims about one superior AI assessor. A tool may align well with one learner profile, but less well with another. In assessment terms, the issue is not only whether the AI score is close to the human score. The more important question is whether the tool follows the human rater's diagnostic logic: which errors are considered important and why, how much weight they are given, and whether the feedback reflects the learner's actual performance rather than a representation shaped solely by the transcript.

| Pooled metric | Mean / overall value | Range |
|---|---|---|
| **Human task mean** | 79.94/100 | 66.62-93.25 |
| **All human task scores** | 57-95/100 | 57-95 |
| **Closest AI configuration MAE** | 7.67 pp | 6.00-9.33 |
| **Second-closest AI configuration MAE** | 9.61 pp | 7.50-11.72 |

*Note. MAE = mean absolute error against the human task-total benchmark; pp = percentage points.*

*Table 4: Pooled AI-human alignment indicators*

### 5.1.4 Assessment-layer divergence and domain instability

The strongest finding of this study is that assessment-layer variation is large, and it is systematic. Under the earlier transcript condition, the mean assessor gap is 25.00 weighted points. Under the

later transcript condition, it rises to 29.00 weighted points. Across conditions, the pooled assessment-layer gap is 27.00 weighted points, with a range of 22.75 to 34.50. Such differences cannot be classified under minor calibration noise. They are large enough to alter feedback, change the interpretation of performance level, and potentially affect pass/fail or progression decisions in borderline cases.

The domain profile sharpens this conclusion. Across the pooled assessment-layer data, *Content* shows a mean gap of 6.63, *Form* 7.75, and *Delivery* 6.00. *Form* is therefore the largest pooled domain gap. This means that the AI assessors diverge especially when judging grammaticality, idiomaticity, register, terminology, source-language interference, and target-language naturalness. These are precisely the features that require expert linguistic and pedagogical judgment, especially in healthcare interpreting, where unclear or unnatural phrasing can impair patient comprehension and therapeutic outcome.

Our collective results suggest two distinct AI severity profiles. One AI assessor behaves as more generous and globally encouraging (Gemini), while the other behaves as stricter and more risk-sensitive (ChatGPT). Generous profile may be useful for formative language, but it can under-penalize serious weaknesses. Stricter profile, on the other hand, may be diagnostically useful, but it can over-penalize transcript-mediated fragmentation. The analysis suggests that AI score should not be treated as a neutral measurement of interpreting quality in any of the cases examined here.

| Condition | Mean total gap | Total-gap range | Content | Form | Delivery | Main implication |
|---|---|---|---|---|---|---|
| **Europa/2025** | 25.00 | 22.75-27.25 | 5.88 | 7.38 | 5.88 | Large assessor divergence appears under the earlier transcript condition. |
| **Notta/2026** | 29.00 | 23.50-34.50 | 7.38 | 8.13 | 6.13 | The assessor gap increases under the later transcript condition. |
| **Assessment layer overall** | 27.00 | 22.75-34.50 | 6.63 | 7.75 | 6.00 | Gemini is consistently more generous; Form is the largest pooled domain gap. |

*Note. MAD = mean absolute criterion difference. Percentages are averaged across the relevant comparison rows.*

*Table 5: Criterion-level agreement indicators*

### 5.1.5 Criterion-level convergence and divergence

The criterion-level indicators show why large total-score differences can emerge even when individual criterion scores appear to be relatively close. In the direct transcription-layer comparison, exact agreement is 53.6%, within-one-point agreement is 94.0%, and MAD is 0.52.

This confirms that transcript source matters, but it also shows that transcription-layer variation is less destabilizing than assessor-layer variation in the pooled data. By contrast, assessment-effect comparisons are much less stable. The pooled exact agreement across assessment-effect rows is only 21.4%, with a range of 7.1% to 31.0%. Within-one-point agreement is higher, at 76.2%, but this should not be overinterpreted. Because the rubric contains 21 criteria and the total is weighted,

repeated one-point differences can accumulate into large total-score gaps. The pooled MAD of 1.04 on a 0-4 criterion scale is therefore practically meaningful.

The all-pipeline exact agreement of 27.9% is misleadingly reassuring unless the layers are separated. It is lifted by the more convergent transcription-layer row. Once the analysis focuses on assessment-effect rows, instability becomes clear. The main validity problem is not simply that transcripts differ; it is that AI assessors convert similar or identical transcript evidence into different judgments.

| Layer / condition | Rows | Exact agreement | Exact range | Within-one | Within-one range | MAD | MAD range | Implication |
|---|---|---|---|---|---|---|---|---|
| **Direct transcription** | 1 | 53.6% | single row | 94.0% | single row | 0.52 | single row | **ASR-source effect is** |
| **n-layer comparison** | | | | | | | | visible but less divergent than assessor effects. |
| **Assessment effect, Europa/ 2025** | 2 | 23.8% | 21.4-26.2% | 77.4% | 75.0-79.8% | 1.01 | 0.96-1.06 | AI assessors differ substantially on the same transcript source. |
| **Assessment effect, Notta/ 2026** | 2 | 19.1% | 7.1-31.0% | 75.0% | 71.4-78.6% | 1.07 | 0.90-1.24 | Agreement remains unstable under the later transcript condition. |
| **Assessment effect overall** | **4** | **21.4%** | **7.1-31.0%** | **76.2%** | **71.4-79.8%** | **1.04** | **0.90-1.24** | Local criterion differences accumulate into large total-score gaps. |

| Layer / condition | Rows | Exact agreement | Exact range | Within-one | Within-one range | MAD | MAD range | Implication |
|---|---|---|---|---|---|---|---|---|
| **n-layer comparison** | | | | | | | | visible but less divergent than assessor effects. |
| **Assessment effect, Europa/ 2025** | 2 | 23.8% | 21.4-26.2% | 77.4% | 75.0-79.8% | 1.01 | 0.96-1.06 | AI assessors differ substantially on the same transcript source. |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Assessment effect, Notta/ 2026** | 2 | 19.1% | 7.1-31.0% | 75.0% | 71.4-78.6% | 1.07 | 0.90-1.24 | Agreement remains unstable under the later transcript condition. |
| **All available pipeline comparisons** | **5** | **27.9%** | **7.1-53.6%** | **79.8%** | **71.4-94.0%** | **0.94** | **0.52-1.24** | The pooled pipeline looks more stable only because the ASR row is more convergent. |

*Note. MAD = mean absolute criterion difference. Percentages are averaged across the relevant comparison rows.*

*Table 6: Assessment-layer gaps by transcript/time condition*

### 5.1.6 Human and AI convergence

The evidence suggests that human and AI assessment converge most strongly when the relevant evidence is textually supported and recoverable. This includes omissions, additions, distortions, contradictory meanings, terminology errors, incomplete rendering of main ideas, and basic syntactic plausibility. In these areas, AI can serve as a useful second reader because it forces the instructor/rater to re-examine whether a global impression is masking local problems.

AI is also useful for making assessments more auditable. A human rater may know that a student performance was weak, but the AI prompt can provide explicit articulation of the reason: whether the problem lies in content transfer, target-language form, or delivery. This can support feedback writing, moderation, and rubric refinement without substituting rater's judgment. After all, the rater-instructor is the only person who has a comprehensive picture of each individual student, their strengths and weaknesses, the conditions under which they take the exam and how they tend to behave and perform under straneous circumstances. The value of AI is therefore not that it replaces human judgment, but that it can make parts of that judgment more explicit, easier to identify and compare.

However, the pooled data also show that convergence at the level of error detection does not guarantee convergence at the level of score. AI systems may identify the same error locus and still assign very different penalties. The central issue is evaluative proportionality: how strongly the error should affect the score given the communicative purpose of the interpreting task. This task should remain the perview of the instructor-rater.

### 5.1.7 Human and AI deviation

Human and AI judgment deviate most clearly in three circumstances. First, they deviate when oral-performance features are central. Delivery features such as pronunciation, intonation, pausing, rhythm, hesitation, false starts, confidence, pitch, stress, and interactional timing cannot be fully inferred from transcript text. Even a high-quality transcript remains only a partial representation of oral performance.

Second, they deviate when transcript fidelity is unstable. A transcript can normalize disfluency, omit hesitation, distort terminology, mis-segment turns, introduce lexical substitutions that were

not present in the original performance or fail to recognize pronunciations and dialectical and/or idiomatic speech produnction. In such cases, the AI may assess a representation of the performance rather than the performance itself. The resulting score may be internally coherent and still invalid.

Third, AI systems deviate when penalty calibration differs. The overall assessment-layer gap of 27.00 weighted points shows that the same transcript evidence can produce markedly different scoring outcomes. One AI assessor may preserve a favorable global judgment because the performance remains broadly intelligible (Gemini); another may penalize local weaknesses more heavily because they are deemed clinically or pedagogically consequential (ChatGPT). The human rater must therefore decide not only whether AI observation is true, but whether its scoring consequence is proportionate not only within the task itself but also across tasks as well as across students.

### 5.1.8 Practical coordination of human and AI assessment

The overall findings support a human-led coordination model rather than a human-versus-AI model. It supports a sequence of actions whereby the instructor first listens to the audio and assigns an independent score before consulting AI output. This preserves the primacy of expert judgment while reducing anchoring effects. The study suggests that AI may enter the workflow only after the human rater has established an initial evaluative position.

The most efficient routine use of AI is not full parallel scoring by multiple tools. It is a targeted discrepancy scan. AI should be asked to flag omissions, additions, contradictions, terminology errors, clinically consequential distortions, and possible opposite meanings. These are areas where transcript-based analysis can be useful and where the instructor can verify the claim against the audio and source material.

A second AI assessor should be reserved for borderline cases, unexpectedly large discrepancies, or high-stakes decisions. Otherwise, multiple AI outputs can generate considerable noise, inflate confidence, create unnecessary cognitive burden, thus resulting in assessment complications. Once the score has been decided by the human rater, AI can be used more safely to help formulate clear, structured, pedagogically useful feedback.

| Workflow stage | Pooled recommendation | Reason | Efficiency implication |
|---|---|---|---|
| **Human first score** | Score from audio before consulting AI. | Prevents anchoring by AI severity or generosity. | Essential. |
| **Transcript verification** | Use one preferred transcript; check high-stakes segments against audio. | Full multi-ASR comparison is research-rich but routine- costly. | Selective verification. |
| **AI discrepancy scan** | Ask AI to flag omissions, additions, distortions, terminology errors and contradictions. | Uses AI for textual-linguistic checking, where it is strongest. | High benefit if prompted tightly. |
| **Second AI only if needed** | Use another AI only for borderline or highly discrepant cases. | Multiple AI outputs can create noise and anchoring pressure. | Selective, not routine. |
| **Human adjudication** | Accept, reject or revise AI observations after evidence checking. | Prevents hallucination and transcript-induced errors from clouding judgment. | Crucial safeguard. |
| **Feedback synthesis** | Use AI to phrase feedback after score is decided. | Captures pedagogical benefit while limiting scoring bias. | Efficient and low risk. |

*Note. This table is procedural rather than averaged; it converts the pooled findings into a routine-use model.*

*Table 7: Human-AI coordination model for routine use*

| Global conclusion | Pooled statistical basis | Practical implication |
|---|---|---|
| **AI-human alignment is profile-dependent.** | Closest-configuration MAE: mean 7.67 pp, range 6.00-9.33. | No single AI tool should be declared globally superior. |
| **Assessment variation exceeds transcript variation.** | Assessment-layer mean gap = 27.00; direct transcription-effect row = 10.75. | Prioritize assessor calibration and prompt control. |
| **Form is the most unstable domain.** | Mean domain gaps: Content 6.63; Form 7.75; Delivery 6.00. | Clarify Form descriptors and anchors. |
| **AI assessors are not interchangeable.** | Assessment-effect exact agreement mean = 21.4%; range 7.1-31.0%. | Use AI for discrepancy detection, not autonomous scoring. |
| **The full workflow is not routine-efficient** | Requires transcription, verification, multiple AI runs, extraction, normalization and adjudication. | Use full pipeline for research; simplify for routine assessment. |

*Table 8: General Conclusions*

### 5.1.9 Time, effort, and feasibility for a single instructor-rater

The collective conclusions make the feasibility issue unavoidable. The full experimental workflow is methodologically valuable but operationally extremely heavy. It requires recording management, anonymization, transcription, transcript checking, advanced AI prompting, score extraction, scale normalization, domain-level comparison, criterion-level comparison, qualitative discrepancy analysis, parallel handling of several hundred files of several layers and combinations of comparing data (over 200 files in our case), as well as final adjudication. For a single instructor working under normal teaching constraints, this is not a routine marking workflow; it is a research and validation workflow.

The benefits are nevertheless real and should not be underestimated. The process exposes hidden assumptions in human scoring, tests the robustness of the rubric, identifies unstable domains, reveals the severity profiles of AI assessors, and helps distinguish transcript-induced artifacts from genuine performance weaknesses. It can also improve feedback by making the reasons for a score more explicit. For research, moderation, high-stakes review, or borderline cases, these benefits may justify the time investment. A severe AI-generated report may also lead the instructor to retrospectively question a score that was, in fact, fair. This is particularly problematic when the human score does not merely assess individual performance, but also situates that performance in relation to the cohort as a whole. In doing so, the instructor's assessment may reveal not only student shortcomings, but also the underlying reasons for them. These reasons may be linked to interpreting instruction itself and to the methodological choices made during testing, including the appropriateness of the assessment tasks and tools. At the same time, a hallucinated AI observation may redirect attention toward an error that does not exist, while a transcript-induced distortion may be misinterpreted as a problem in the student's actual performance.

The danger is not only that AI may be wrong, but that it may be persuasively wrong. For this reason, the benefits outweigh the costs only when the workflow is proportionate to the purpose. Full pooled or multi-layer analysis is justified for research, tool validation, assessment design, and difficult moderation cases. Routine classroom use should be simplified to a human-first, AI-second model centered on discrepancy detection and feedback support.

## 6 Limitations

Several limitations must be acknowledged in association with our experiment. First, the study is based on a small-scale exploratory corpus drawn from a single cohort of six MA students in one

healthcare interpreting course. Although the corpus is ecologically valid because it consists of authentic examination performances, its findings cannot be generalized statistically to all interpreting students, language pairs, institutions or assessment contexts. Consequently, the purpose of the study is not to allow for broad generalization, but for theoretically informed and empirically grounded exploration of how AI-assisted assessment behaves under realistic pedagogical conditions.

Second, the corpus is limited to the English-French (Canadian) language pair and to healthcare-related tasks focused on one topic, namely the gastrointestinal system. The findings may not transfer directly to other language combinations and/or regional varieties, especially lower-resource languages, languages with different morphosyntactic profiles, or interpreting contexts involving different registers, discourse genres, thematic topics or institutional constraints. The performance of both ASR systems and AI assessors may vary significantly according to language direction, accent, terminology density, recording quality, and interactional complexity.

Third, transcript fidelity remains a major methodological limitation. AI assessment in this study depends substantially on machine-generated transcripts, yet the analysis shows that transcripts do not simply reproduce performance; they reshape it. Disfluencies, pauses, false starts, self-corrections, pronunciation problems, turn-taking signals, and prosodic features may be omitted, normalized, or inadequately represented. As a result, some AI judgments may reflect the transcript's representation of the performance rather than the performance itself. This limitation is especially consequential for *Delivery*, and *Form* but it may also affect *Content* whenever transcription errors distort terminology, speaker turns, qualifiers, hedges, or meaning-bearing lexical items.

Fourth, the transcription tools themselves presented uneven and tool-specific limitations. Some tools were unable to produce strict verbatim transcripts under the required conditions, while others normalized speech in ways that reduced the visibility of delivery-related phenomena. Notta.ai proved more useful in this particular workflow, but even its output required human checking. Europa and Notta displayed different error profiles, and these differences matter because substitution-heavy, deletion-heavy, and insertion-heavy transcripts do not affect downstream assessment in the same way. For these reasons, the study cannot identify one transcription system as globally superior; it can only describe vulnerability profiles under the specific conditions tested and practical ways to overcome them where applicable.

Fifth, AI-generated assessment remains prompt-, tool-, and version-dependent. The study compares outputs produced at different points in time, but AI systems evolve rapidly and often without full transparency regarding model architecture, training data, system behavior, or interface-level changes. This limits replicability. A prompt that produces one type of assessment at one point in time may produce a different assessment later, even when the same files and rubrics are used.

Sixth, the human assessor's role is both a factor of strengths and weaknesses. The instructor-rater provides expert contextual judgment, direct access to the pedagogical setting, and the ability to interpret performance in relation to course expectations and cohort specificities. However, as in all human assessment, the process remains vulnerable to subjectivity, prior knowledge of students, fatigue, memory effects, and possible confirmation bias. The additional human reassessment after exposure to AI outputs is useful for reflection, but it may also be affected by AI anchoring, whereby machine-generated observations influence the rater's subsequent interpretation. This reinforces the

need for procedural sequencing in future human-AI assessment models, with initial human scoring ideally completed before AI outputs are consulted.

Seventh, the weighted analytic grid itself is still under validation. Its *Content*, *Form*, and *Delivery* structure proved useful for identifying domain-specific patterns, but the findings also suggest that some descriptors may require refinement, especially where criteria overlap or where small criterion-level differences accumulate into large total-score gaps. Future work should examine whether the weighting system, descriptor specificity, and domain boundaries require recalibration in light of empirical evidence from both human and AI-assisted scoring.

Finally, the study raises ethical and data-governance limitations. AI-assisted transcription and assessment were conducted through personal paid subscriptions rather than institutionally governed environments. Although participant consent was obtained, files were anonymized, privacy controls were used where available, and transcripts were human-checked, consumer-grade platforms provide less institutional control than environments governed by formal organizational agreements. Moreover, voice recordings may remain identifying even when transcripts are anonymized, and ASR performance may vary across speaker profiles. These residual privacy, fairness, and validity risks cannot be fully eliminated. They were mitigated in the present study, but they remain part of the ethical boundary conditions of AI-assisted research with student audio data. In light of these limitations, the findings should be interpreted as evidence for cautious, bounded, and pedagogically supervised use of AI in interpreting assessment.

## Conclusion

This study sets out to examine whether readily available AI tools can support the assessment of student interpreting performance in examination settings, and whether such support can be considered valid, reliable, pedagogically meaningful, and practically feasible for instructors working without specialized computational infrastructure and/or expert knowledge. As announced earlier in this paper, the analysis combined quantitative comparison of scoring patterns with qualitative examination of recurrent convergences and divergences between human and AI-generated evaluations. It also considered the longitudinal dimension of AI assessment by comparing outputs produced at different moments in time, while treating transcript fidelity as a central methodological variable rather than as a neutral preprocessing step.

The synthesis confirms that AI-assisted assessment can make interpreting assessment more explicit, more auditable, and more diagnostically structured. Its principal value lies in its capacity to render assessment decisions more visible: it can identify candidate omissions, additions, distortions, terminological problems, inconsistencies in meaning transfer, problems of coherence, and syntactic or lexical weaknesses that may otherwise remain implicit in the instructor's overall judgment. In this respect, AI can function as a useful diagnostic support tool, especially when assessment focuses on textually recoverable aspects of performance. This is particularly evident in the domains of *Content* and, to a lesser extent, *Form*, where the comparison between student output, source text, and assessment grid allows AI systems to generate structured observations about transfer accuracy, completeness, lexical choice, terminology, grammar, cohesion, and plausibility of formulation.

At the same time, the findings also confirm one of the central claims anticipated in this study: AI assessment remains significantly less reliable for *Delivery*. Delivery is not reducible to what appears in the transcript. It depends on pronunciation, intonation, stress, rhythm, speed, pausing,

pitch, loudness, hesitation patterns, turn-taking cues, pragmatic force, speaker attitude, emotional colouring, and the broader acoustic and interactional profile of the performance. These features are either absent, reduced, or normalized in most machine-generated transcripts. Consequently, human listening remains indispensable. Where AI appears to assess delivery, it often does so through indirect textual proxies rather than through full access to the oral performance as experienced by a human rater. This is a major validity boundary for AI-assisted assessment of interpreting.

The central empirical pattern emerging from the analysis is therefore not one of simple AI-human agreement or disagreement. Rather, AI-human alignment is profile-dependent. It varies by student, task type, language direction, language variety, transcription source, assessment tool, rubric domain, and criterion. The study shows that assessment-layer divergence is larger than direct transcription-layer divergence: even when transcript differences appear relatively contained, AI-generated evaluative judgments may diverge more substantially from human assessment. This indicates that error does not enter the assessment process only through transcription; it also enters through interpretation of the rubric, weighting of evidence, inferential reasoning, and the AI system's tendency to over- or under-penalize particular features of performance.

The results also show that *Form* is the most unstable domain. This instability is analytically important because *Form* lies between the more textually recoverable domain of *Content* and the more acoustically dependent domain of *Delivery*. It includes grammar, syntax, idiomaticity, language variety, terminology, coherence, cohesion, and quality of expression. These criteria are partly visible in transcripts, but they are also sensitive to the assessor's expectations regarding acceptable oral formulation, interpreter-mediated reformulation, register, communicative adequacy, and the difference between written-like correctness and spoken interpreting performance. AI tools may therefore reward linguistic polish or penalize oral reformulation patterns in ways that do not always correspond to expert interpreting pedagogy.

Another important finding is that AI assessors are not interchangeable. Different tools do not merely produce different wordings of the same judgment; they may produce different evaluative profiles. They may assign different relative importance to omissions, distortions, fluency problems, lexical choice, grammatical accuracy, or delivery-related phenomena. This is not in itself detrimental to the process: human assessment may also differ from one rater to another. AI tools may also differ in their vulnerability to transcript-mediated bias, hallucination-like inference, overconfidence, excessive strictness, or excessive generosity. This confirms that AI-assisted assessment should not be treated as a generic technological category. Each tool has its own vulnerability profile, and each output requires contextual verification by a competent human assessor.

The study further demonstrates that small criterion-level discrepancies can accumulate into substantial total-score differences. This is especially important in a weighted analytic grid containing multiple criteria across *Content*, *Form*, and *Delivery*. A one-point difference on several criteria may appear minor when examined locally, but it can produce a meaningful shift in the total score, particularly when *Content* is weighted more heavily. For this reason, within-one-point agreement should not be overinterpreted as evidence of practical equivalence. In high-stakes assessment, repeated small divergences may affect grades, progression, feedback, and students' perception of fairness.

Taken together, these findings argue strongly against autonomous AI scoring. They do not support replacing the instructor-rater with an AI tool, nor do they support treating AI-generated

scores as independent assessment outcomes. Instead, they support a carefully bounded model of human-AI coordination. In such a model, the instructor should score first. AI should then be used selectively to verify transcript-based evidence, identify possible discrepancies, detect overlooked omissions or distortions, and generate diagnostic prompts for further human reflection. A second AI tool may be useful in borderline, complex, or high-stakes cases, but not as a routine substitute for or verification of expert judgment. Final adjudicative authority must remain with the human assessor.

The practical implication is therefore not to replace humans with AI but to implement a controlled support procedure. When used appropriately, AI can help instructors ask better questions about a score: Was a key qualifier omitted? Was a medical term distorted and, if yes, what was the main reason for such a distorsion? Was a syntactic problem serious enough to impede comprehension? Did the student preserve the main idea but lose pragmatic nuance? Did the transcript itself misrepresent the student's performance? In this way, AI can contribute to transparency, auditability, feedback quality, and rater self-monitoring. Used uncritically, however, it can cloud the assessor's judgment, amplify transcript errors, introduce hallucinated evidence, normalize bias, and create a false impression of objectivity in a process that remains deeply contextual, ethical, and pedagogical.

Against this backgrop we believe that the broader contribution of the study is twofold. First, it shows that mainstream AI tools can have meaningful pedagogical value in interpreter education when they are embedded in a human-in-the-loop assessment architecture. This is, for example, the case, of Notta.ai, which has incorporated features that allow the instructor-rater to insert comments, observations or corrections to the produced transcript. Advancements in ASR technology are significant but still require substantial human intervention and verification. Second, it demonstrates that the validity of AI-assisted interpreting assessment depends on the integrity of the entire representation chain: source text, source audio, student performance, transcript, rubric, prompt, AI output, and human adjudication. AI can support assessment only when this chain is made explicit, checked, and critically interpreted at each stage. The more opaque or unstable the chain becomes, the greater the risk that AI assessment will evaluate not the student's interpreting performance, but a distorted textual representation of it.

## References


Alon, L., & Levkovich, I. (2026). Trusting the black box: Adapting a multidimensional measure of trust in generative AI. Computers in Human Behavior: Artificial Humans, 8, 100295. https://doi.org/10.1016/j.chbah.2026.100295

Bäckström, T. (2025). Privacy in speech technology. *Proceedings of the IEEE, 113*(7), 668–692. Doi: 10.1109/JPROC.2025.3632102

Barnett, J. (2023). The ethical implications of generative audio models: A systematic literature review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. Doi: 10.1145/3600211.3604686

Bélisle-Pipon, J.-C., Powell, M., English, R., Malo, M.-F., Ravitsky, V., Bridge2AI–Voice Consortium, & Bensoussan, Y. (2024). Stakeholder perspectives on ethical and trustworthy voice AI in health care. *Digital Health, 10*. Doi:10.1177/20552076241260407

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587–604. Doi: 10.1162/tacl_a_00041

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Doi: 10.1145/3442188.3445922

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, *3*, Article 22. https://doi.org/10.3389/feduc.2018.00022

Da Silva, J. (2021). Producing "good enough" automated transcripts securely: Extending Bokhove and Downey (2018) to address security concerns. *Methodological Innovations, 14*(1). https://doi.org/10.1177/2059799120987766

Eftekhari, H. (2024). Transcribing in the digital age: Qualitative research practice utilizing intelligent speech recognition technology. *European Journal of Cardiovascular Nursing, 23*(5), 553–560. doi:10.1093/eurjcn/zvae013

Feng, S., Halpern, B. M., Kudina, O., & Scharenborg, O. (2023). Towards inclusive automatic speech recognition. *Computer Speech & Language, 84*, 101567. https://doi.org/10.1016/j.csl.2023.101567

Google. (n.d.). *Gemini API documentation and product materials*. Accessed: 15 February 2026.

Han, C., & Lu, X. (2021). Interpreting quality assessment re-imagined: The synergy between human and *machine scoring*. *Interpreting and Society, 1*(1), 70–90. https://doi.org/10.1177/27523810211033670 (ResearchGate)

Han, C., & Lu, X. (2025). *Beyond BLEU: Repurposing neural-based metrics to assess interlingual interpreting in tertiary-level language learning settings*. *Research Methods in Applied Linguistics, 4*(1), 100184. https://doi.org/10.1016/j.rmal.2025.100184

Han, C., Lu, X., & Chen, S. (2025). *Modeling rater judgments of interpreting quality: Ordinal logistic regression using neural-based evaluation metrics, acoustic fluency measures, and computational linguistic indices*. *Research Methods in Applied Linguistics, 4*(1), 100194. https://doi.org/10.1016/j.rmal.2025.100194

Herdiyanti, A. (2024). The use of automatic AI-based notes and transcription services in qualitative research: Ethical and methodological concerns. In *Proceedings of the ALISE Annual Conference*. Doi: 10.21900/j.alise.2024.1717

Jiang, Z., & Zhang, Z. (2025). *From black box to transparency: Enhancing automated interpreting assessment with explainable AI in college classrooms*. *Research Methods in Applied Linguistics, 4*(3), 100237. https://doi.org/10.1016/j.rmal.2025.100237

Jönsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*, 7684–7689. https://doi.org/10.1073/pnas.1915768117

Kröger, J.L., Lutz, O.HM., Raschke, P. (2020). Privacy Implications of Voice and Speech Analysis – Information Disclosure by Inference. In: Friedewald, M., Önen, M., Lievens, E., Krenn, S., Fricker, S. (eds) Privacy and Identity Management. Data for Better Living: AI and Privacy. Privacy and Identity 2019. IFIP Advances in Information and Communication Technology(), vol 576. Springer, Cham. https://doi.org/10.1007/978-3-030-42504-3_16

Leschanowsky, A., Rusti, C., Quinlan, C., Pnacek, M., Gorce, L., & Hutiri, W. (2025). A Data Perspective on Ethical Challenges in Voice Biometrics Research. IEEE Transactions on Biometrics, Behavior, and Identity Science, 7(1), 118-131. https://doi.org/10.1109/TBIOM.2024.3446846

Macháček, D., Bojar, O., & Dabre, R. (2023). *MT metrics correlate with human ratings of simultaneous speech translation*. In E. Salesky, M. Federico, & M. Carpuat (Eds.), *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)* (pp. 169–179). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.iwslt-1.12

McCowan, I. A., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., & Bourlard, H. (2005). *On the use of information retrieval measures for speech recognition evaluation* (IDIAP Research Report No. 04-73). IDIAP Research Institute.

McMullin, C. (2023). Transcription and qualitative methods: Implications for third sector research. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations, 34*(1), 140–153. https://doi.org/10.1007/s11266-021-00400-3

Morris, A. C., Maier, V., & Green, P. (2004). From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. *Proceedings of Interspeech 2004*, 2765–2768. https://doi.org/10.21437/Interspeech.2004-668

Nautsch, A., Jasserand, C., Kindt, E., Todisco, M., Trancoso, I., & Evans, N. (2019). The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding. In *Proceedings of Interspeech 2019* (pp. 3695–3699). https://doi.org/10.21437/Interspeech.2019-2647

Nautsch, A., Jiménez, A., Treiber, A., Kolberg, J., Jasserand, C., Kindt, E., Delgado, H., Todisco, M., Hmani, M. A., Mtibaa, A., Abdelraheem, M. A., Abad, A., Teixeira, F., Gomez-Barrero, M., Petrovska-Delacrétaz, D., Chollet, G., Evans, N., Schneider, T., Bonastre, J.-F., Raj, B., Trancoso, I., & Busch, C. (2019b). Preserving privacy in speaker and speech characterisation. *Computer Speech & Language, 58*, 441–480. Doi: 10.1016/j.csl.2019.06.001

Notta. (n.d.). *Product and privacy documentation*. Accessed: 5 February 2026

OpenAI. (n.d.). *Speech-to-text and ChatGPT product documentation*. Accessed: 21 January 2026

Panadero, E., & Jönsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129–144. https://doi.org/10.1016/j.edurev.2013.01.002

Pöchhacker, F. (2001). Quality assessment in conference and community interpreting. *Meta, 46*(2), 410–425. https://doi.org/10.7202/003847ar

Samuel, G., & Wassenaar, D. (2025). Joint editorial: Informed consent and AI transcription of qualitative data. *Journal of Empirical Research on Human Research Ethics, 20*(1–2), 3–5. https://doi.org/10.1177/15562646241296712

Shafiei, S. (2024). A proposed analytic rubric for consecutive interpreting assessment: Implications for similar contexts. *Language Testing in Asia, 14*, Article 13. https://doi.org/10.1186/s40468-024-00278-0

Srivastava, B. M. L., Maouche, M., Sahidullah, M., Vincent, E., Bellet, A., Tommasi, M., Tomashenko, N., Wang, X., & Yamagishi, J. (2022). Privacy and utility of x-vector based speaker anonymization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30*, 2383–2395. doi:10.1109/TASLP.2022.3190741

Stewart, C., Vogler, N., Hu, J., Boyd-Graber, J., & Neubig, G. (2018). *Automatic estimation of simultaneous interpreter performance*. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 662–666). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2105

Tomashenko, N., Wang, X., Vincent, E., Patino, J., Srivastava, B. M. L., Noé, P.-G., Nautsch, A., Evans, N., Yamagishi, J., O'Brien, B., Chanclu, A., Bonastre, J.-F., Todisco, M., & Maouche, M. (2022). The VoicePrivacy 2020 Challenge: Results and findings. *Computer Speech & Language, 74*, Article 101362. Doi: 10.1016/j.csl.2022.101362

Wang, X., & Wang, B. (2024). *Identifying fluency parameters for a machine-learning-based automated interpreting assessment system*. *Perspectives, 32*(2), 278–294. https://doi.org/10.1080/0907676X.2022.2133618

Wang, X., & Wang, B. (2025). *Advancing automatic assessment of target-language quality in interpreter training with large language models: Insights from explainable AI*. *The Interpreter and Translator Trainer, 19*(3–4), 465–485. https://doi.org/10.1080/1750399X.2025.2533015

Wang, X., & Yuan, L. (2023). *Machine-learning based automatic assessment of communication in interpreting*. *Frontiers in Communication, 8*, 1047753. https://doi.org/10.3389/fcomm.2023.1047753

## Annex 1 – Student interpreting assessment grid

| Assessment Categories & Criteria | Level of Effectiveness | | | | |
|---|---|---|---|---|---|
| **Note to User:**<br>1. There are three major categories: Content, Form and Delivery<br>2. Arrangement of Criteria is done in the latter's order of importance (weight)<br>3. Each Category contains 7 Criteria | **4 COMPLETE**<br>***ALL*** characteristics present | **3 EXTENSIVE**<br>***MOST*** characteristics present | **2 MODERATE**<br>***SOME*** characteristics present | **1 LIMITED**<br>***FEW*** characteristics present | **0 ZERO**<br>***NO*** characteristics Present |

| | **I. CONTENT** | **COMMENTS** |
|---|---|---|
| 1 | ( ) No opposite meaning | |
| 2 | ( ) Accurate rendition of main ideas (no omissions) | |
| 3 | ( ) No unjustified change in meaning | |
| 4 | ( ) Logical cohesion & coherence | |
| 5 | ( ) Completeness of information (general) | |
| 6 | ( ) Completeness of information (names & numbers) | |
| 7 | ( ) No unjustified additions | |
| | Subtotal: | |
| | **II. FORM** | |
| 1 | ( ) No incomplete sentences | |
| 2 | ( ) Natural/Idiomatic target-language expressions | |
| 3 | ( ) Unambiguous & clear diction | |
| 4 | ( ) Appropriate register and speech level (genre) | |
| 5 | ( ) Little source-language interference | |
| 6 | ( ) Correct terminology | |
| 7 | ( ) Grammatical correctness | |
| | Subtotal: | |
| | **III. DELIVERY** | |
| 1 | ( ) Fluency of delivery (general concept/impression) | |
| 2 | ( ) No significant repairs or backtracking (self-correction) | |
| 3 | ( ) Impression of confidence | |
| 4 | ( ) Few fillers, hesitations and pauses | |
| 5 | ( ) Lively intonation, proper stress and good pronunciation | |
| 6 | ( ) Finishing interpretation within time limit | |
| 7 | ( ) No slips of tongue | |
| | Subtotal: | |
| | **TOTAL: /112** | **Total = content score x 2 + form score x 1 + delivery score x 1** |

## Annex 2 – Prompts

### For bidirectional dialogue

**Role:** You are an expert healthcare interpreter instructor and a meticulous transcriber.
**Task:** Listen to the provided audio file of a student's medical interpreting roleplay. Produce a **strict verbatim transcript** of the entire audio file to be used for grading and assessment.
**Transcription Guidelines:** You must capture every single sound and utterance exactly as it occurs in the recording. Do not clean up the grammar, and do not remove any disfluencies. Strictly adhere to the following formatting rules:

- **Fillers & Hesitations:** Include all vocalized pauses exactly as they sound (e.g., *um, uh, ah, er, mm-hmm*).
- **Stutters & False Starts:** Transcribe stutters using hyphens (e.g., *I- I- I think that...* or *the p-p-patient*). If a speaker starts a sentence, stops, and changes direction, capture the exact break (e.g., *We need to— let's look at the chart.*).
- **Pauses:** Indicate noticeable silences using brackets. Use [pause] for standard breaks (1-3 seconds) and [long pause] for anything longer than 3 seconds.
- **Mispronunciations:** Type the word exactly as the student pronounced it phonetically, followed by the intended word in brackets (e.g., *They need an endo-scoppy [endoscopy]*).
- **Non-Verbal Sounds:** Note important non-verbal sounds that affect communication, such as [sigh], [clears throat], or [laughs].
- **Distortions & Inaudible Speech:** If the audio drops out or the student mumbles beyond comprehension, mark it as [inaudible] or [audio distortion].
- **Speaker Labels:** Clearly identify speakers (e.g., **Doctor:**, **Patient:**, **Student Interpreter:**). If the student switches languages, note the language being spoken in parentheses next to their name (e.g., **Student Interpreter (French):**).

### For consecutive interpreting

**Role:** You are an expert healthcare interpreter instructor and a meticulous transcriber.
**Task:** Listen to the provided audio file of a student's medical interpreting consecutive rendition from English into French/French into English. Produce a **strict verbatim transcript** of the entire audio file to be used for grading and assessment.
**Transcription Guidelines:** You must capture every single sound and utterance exactly as it occurs in the recording. Do not clean up the grammar, and do not remove any disfluencies. Strictly adhere to the following formatting rules:

- **Fillers & Hesitations:** Include all vocalized pauses exactly as they sound (e.g., *um, uh, ah, er, mm-hmm*).

- **Stutters & False Starts:** Transcribe stutters using hyphens (e.g., *I- I- I think that...* or *the p-p-patient*). If a speaker starts a sentence, stops, and changes direction, capture the exact break (e.g., *We need to— let's look at the chart.*).
- **Pauses:** Indicate noticeable silences using brackets. Use [pause] for standard breaks (1-3 seconds) and [long pause] for anything longer than 3 seconds.
- **Mispronunciations:** Type the word exactly as the student pronounced it phonetically, followed by the intended word in brackets (e.g., *They need an endo-scoppy [endoscopy]*).
- **Non-Verbal Sounds:** Note important non-verbal sounds that affect communication, such as [sigh], [clears throat], or [laughs].
- **Distortions & Inaudible Speech:** If the audio drops out or the student mumbles beyond comprehension, mark it as [inaudible] or [audio distortion].

**For Student Assessment**

You are an expert instructor in healthcare interpreting and a highly experienced performance assessor. Your task is to evaluate a student's medical interpreting performance for the assigned task.
You will be provided with:

1. the audio recording of the student's interpreted performance,
2. the transcript of the student's performance,
3. the audio recording of the original source speech or dialogue that the student was asked to interpret,
4. the transcript of the original source speech or dialogue, and
5. the interpreting evaluation grid.

Assess the student's performance by comparing the student's rendition against the original source audio and transcript, while also using the student transcript as supporting evidence. Base your evaluation strictly on the attached rubric, which contains three domains — **Content, Form, and Delivery** — with **seven criteria under each domain**, rated on a **0–4 scale**. The grid defines the scale as follows: **4 = Complete, 3 = Extensive, 2 = Moderate, 1 = Limited, 0 = Zero**. The total score is calculated as: **Content subtotal × 2 + Form subtotal × 1 + Delivery subtotal × 1**.
Your assessment must be exhaustive, precise, and evidence-based.
**Instructions for assessment**
Evaluate **every sub-criterion individually** under the three rubric domains.
**I. Content**
Assess the student's performance on the following:

1. No opposite meaning
2. Accurate rendition of main ideas (no omissions)

3. No unjustified change in meaning
4. Logical cohesion and coherence
5. Completeness of information (general)
6. Completeness of information (names and numbers)
7. No unjustified additions

For each criterion:

- assign a score from 0 to 4,
- explain the score in detail,
- identify any critical omissions, distortions, contradictions, additions, or losses of medically relevant information,
- state whether the issue is minor, major, or critical in relation to patient safety, clinical accuracy, or communicative adequacy.

**II. Form**

Assess the student's performance on the following:

1. No incomplete sentences
2. Natural/idiomatic target-language expressions
3. Unambiguous and clear diction
4. Appropriate register and speech level (genre)
5. Little source-language interference
6. Correct terminology
7. Grammatical correctness

For each criterion:

- assign a score from 0 to 4,
- explain the score in detail,
- comment on linguistic quality, clarity, appropriateness to the medical setting, and target-language acceptability,
- identify any terminology errors, grammatical problems, awkward phrasing, calques, ambiguity, or register mismatches.

**III. Delivery**

Assess the student's performance on the following:

1. Fluency of delivery (general impression)

2. No significant repairs or backtracking (self-correction)
3. Impression of confidence
4. Few fillers, hesitations, and pauses
5. Lively intonation, proper stress, and good pronunciation
6. Finishing interpretation within the time limit
7. No slips of the tongue

For each criterion:

- assign a score from 0 to 4,
- explain the score in detail,
- comment on speech flow, hesitations, self-repairs, pauses, prosody, pronunciation, confidence, and overall presentational control,
- identify whether delivery issues interfere with comprehension, professional credibility, or communicative effectiveness.

**Important requirements**

- Assess **every single aspect** of the student's performance.
- Discuss **all 21 criteria** explicitly. Do not skip any sub-element.
- Ground your evaluation in the evidence provided by the audio and transcripts.
- Give special weight to errors affecting **medical meaning, patient safety, names, numbers, dosage, symptoms, procedures, chronology, negation, and clinical instructions**.
- Distinguish clearly between:
    - minor deviations that do not materially affect meaning,
    - major deviations that reduce accuracy or clarity,
    - critical deviations that could compromise medical understanding, patient safety, or the communicative purpose of the encounter.
- Do not rely only on the transcript if the audio reveals hesitations, pronunciation issues, false starts, repairs, omissions, or delivery problems not visible in the written text.
- Where relevant, comment on whether transcript-based evidence may underrepresent or overrepresent performance features that are only audible in the recording.

**Required output format**
Use the following structure:
**1. Task overview**

- Briefly identify the interpreting task, language direction, and communicative setting.

**2. Domain-by-domain assessment**
**A. Content**

- Criterion 1: [score/4] + detailed justification
- Criterion 2: [score/4] + detailed justification
- …
- Criterion 7: [score/4] + detailed justification
- **Content subtotal: X/28**

**B. Form**

- Criterion 1: [score/4] + detailed justification
- Criterion 2: [score/4] + detailed justification
- …
- Criterion 7: [score/4] + detailed justification
- **Form subtotal: X/28**

**C. Delivery**

- Criterion 1: [score/4] + detailed justification
- Criterion 2: [score/4] + detailed justification
- …
- Criterion 7: [score/4] + detailed justification
- **Delivery subtotal: X/28**

**3. Score calculation**

- Content subtotal = X/28
- Form subtotal = X/28
- Delivery subtotal = X/28
- **Weighted total = (Content × 2) + Form + Delivery = X/112**

**4. Overall evaluative judgement**
Provide a synthetic overall assessment explaining:

- whether the interpretation is satisfactory and functional for the communicative purpose,

- whether the rendition successfully conveys critical medical information,
- whether any deviations are minor, major, or critical,
- whether the performance would be acceptable in a real healthcare interaction,
- the student's main strengths,
- the student's main weaknesses,
- the most urgent priorities for improvement.

**5. Final verdict**
Conclude with one concise paragraph stating whether the performance is:

- highly effective,
- generally effective,
- moderately effective,
- limited,
- or ineffective,

and justify that judgement in relation to the rubric and the communicative demands of medical interpreting.

**For analysis**

Read my article first and use it as the governing analytical framework. Then analyze the 8 attached assessment files through a mixed-methods comparative lens, combining quantitative score comparison with qualitative criterion-based discrepancy analysis. Examine alignment between the two AI tools represented here for assessment across task, domain, and criterion, with particular attention to Content, Form, and Delivery. Identify patterns of convergence, divergence, hallucination, inconsistency, omission, distortion, transcript-mediated bias, and over- or under-penalization. Extract and compare total, domain-level, and criterion-level scores; report numerical differences, agreement rates, averages, ranges, and other relevant indicators. Determine which AI tool is closest to human assessment, in which domains, and by how much. Assess each tool's reliability, consistency, and vulnerability profile, and provide a nuanced final judgment as to whether one tool is more dependable overall or only for specific evaluative purposes. Base every conclusion on explicit evidence from the files and provide numbers throughout. Create charts and quantitative data in Excel for further extraction if necessary. Calibrate your analysis bearing in mind that Europa and Notta are transcription tools and Gemini and ChatGPT are used as assessment tools whose evaluation depends on the accuracy or not of the transcription tools. In light of this information, redo the task.